Boosting Store Sales Through Machine Learning-Informed Promotional Decisions

Yue Qiu[*], Wenbin Wang[‡], Tian Xie[†], Jun Yu[‡], Xinyu Zhang[§]

October 20, 2023

Abstract

The sales of fashion products are influenced by uncertain and heterogeneous demands, necessitating predictive analytics to consider multiple explanatory variables and address the challenge of model uncertainty, which has been overlooked in prior research. To illustrate our solution, we first propose a novel forecasting estimator, which is characterized by provable optimal weighted forecasts and a collection of sub-model forecasts with various model specifications. We then validate our method empirically with store-level sales observations of a well-known international footwear brand, as an example of how a retailer can enhance its sales forecasts and improve promotion decisions after controlling model uncertainty. In a predictive analysis, the results show that controlling for model uncertainty between various predictors and store sales can produce more accurate forecasts of sales. With our proposed estimator, we also demonstrate the heterogeneity of promotion strategy importance for store with high and low previous sales. The additional analysis reveals that combo promotions have the most significant impact, and suggests adjusting the frequencies of gift and combo promotions to boost store sales. However, caution is advised when implementing both gift and combo promotions together to mitigate cannibalization effects.

Keywords: machine learning, model uncertainty, model averaging, fashion sales forecasting, promotion evaluation

[*]Finance School, Shanghai University of International Business and Economics, Shanghai, China.
[†]College of Business, Shanghai University of Finance and Economics, Shanghai, China.
[‡]School of Economics and Lee Kong Chian School of Business, Singapore Management University.
[§]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

1 Introduction

Due to short product life cycles and long manufacturing lead times, a sound forecasting strategy can truly add value to the fashion industry by improving inventory planning and helping marketing teams tailor effective promotional activities. This is especially relevant for store sales propelled by the endeavors of sales associates. However, practitioners in the industry often grapple with uncertain and heterogeneous demands, which makes sales forecasting and promotion evaluation particularly challenging. As highlighted by Liu et al. (2013), fashion product sales—including apparel, shoes, and beauty items—are heavily influenced by various factors, such as seasonality, fashion trends, weather, sales promotions, and macroeconomic conditions. Therefore, predictive analytics must consider many and varied explanatory variables, which inevitably leads to the issue of model uncertainty.

There exists a number of research studies on sales prediction for the fashion products.[1] Conventional approaches involve statistical methods such as ARIMA, moving average, weighted average, and exponential smoothing (Box et al., 2015). To capture highly irregular patterns in sales data, artificial intelligence (AI) methods like fuzzy models, artificial neural networks (ANN), evolutionary neural networks (ENN), and extreme learning machines (ELM) have been introduced. These AI techniques have demonstrated higher accuracy than statistical models (Frank et al., 2003; Au et al., 2008; Sun et al., 2008), albeit at the expense of increased computational time and resource requirements. Recently, many researchers have discovered the advantages of utilizing hybrid methods, which combine AI schemes with statistical models or grey models, in order to strike a balance between efficiency and effectiveness in forecasting tasks (Wong and Guo, 2010; Ni and Fan, 2011; Yesil et al., 2012). Although prior work has extensively explored complex and nonlinear relationships between series, none have explicitly addressed the issue of model specification uncertainty.

In this paper, we propose a new type of forecasting approach, which combines off-the-shelf machine learning algorithms with provable optimal weighted forecasts. In particular, machine learning techniques can be employed to generate a collection of forecasts based on various submodels, which are constructed through the full permutation of all potential predictors. We then

[1] The survey paper by Liu et al. (2013) reviews the evolution of analytical methods over the last 15 years, spanning statistical techniques, artificial intelligence models, and hybrid methods. In another review article, Beheshti-Kashi et al. (2015) extend their survey to new product forecasting and the predictive value of user-generated contents, in addition to the topic of fashion product sales.

create a weighted average of these forecasts to predict the response variable, where the weights are selected to minimize the Mallows-type criteria, grounded in the frequentist model averaging literature. To prove the optimality of model weights, it is only necessary for the predictions stemming from any input vector to be a weighted average of observations in response variables. These weights depend on the input vector and the training set, possibly involving a nonlinear relationship. As a result, our framework is sufficiently versatile to accommodate many widely-used machine learning techniques. From the above discussion, it is evident that our approach effectively mitigates the issue of model uncertainty while preserving the nonlinearity between data.

Our approach sheds new light on the forecast combination literature, which was pioneered by Barnard (1963), Reid (1968), and Bates and Granger (1969). In these works, forecasts from sub-models were primarily generated in a parametric setting.[2] In contrast, our proposed hybrid approach addresses model specification uncertainty by leveraging a comprehensive set of sub-model predictions possibly generated through machine learning techniques. Additionally, our approach contributes to the ensemble method from the machine learning literature. Traditionally, an ensemble learning approach aims to enhance predictive accuracy through flexible aggregation schemes that group forecasts from candidate learning algorithms.[3] To the best of our knowledge, there is limited discussion on the statistical properties of related grouping techniques in the literature. Drawing from the model averaging literature, our framework can provide a provably optimal combination scheme for machine learning forecasts under reasonable technical assumptions.

In light of the prevalent use of least squares support vector regression (LSSVR) in business practices, we proceed to assess the effectiveness of our proposed hybrid approach in conjunction with LSSVR. Through Monte Carlo exercises, we initially demonstrate that our proposed hybrid approach consistently outperforms other competitive estimators, particularly in cases involving heteroskedasticity and smaller sample sizes. Utilizing actual weekly data from stores across China of a renowned footwear brand, we aim to examine sales forecasting and evaluate promotion strategies. Our analysis is primarily centered on two sales responses - the weekly count of effective cus-

[2]Excellent reviews of forecast combination techniques can be found in Clemen (1989), Hoeting et al. (1999), Timmermann (2006), and Elliott and Timmermann (2016). Numerous successful applications of forecast combinations in economics and finance have been documented in the literature; see, for instance, Rapach et al. (2010), Elliott et al. (2013), and Genre et al. (2013).

[3]See Sagi and Rokach (2018) and Dong et al. (2020) for recent literature reviews. Notably, to address potential misspecifications in the consumer choice model, Feng et al. (2022) introduced a novel operational data analytics framework to estimate a generalized consumer choice model using data.

tomers and weekly sales revenues. We try to explore the predictors that significantly impact sales forecasts for the upcoming week, which include previous sales records, one-week-ahead district-level weather forecasts, and the weekly frequencies of implementing three promotion strategies, namely gift, combo, and discount promotions. In the predictive analysis with alternative forecasting approaches, we evaluate 15 candidate estimators, ranging from linear estimators, recursive partitioning estimators, support vector regressions, to our hybrid approaches (also referred to as averaging LSSVR in the sequel).

Our results first underscore the value of forecasting methods that can address both model uncertainty and nonlinearity, especially within the context of statistical learning algorithms. With smaller sample sizes, our proposed Mallows averaging LSSVR manifests the best performance for the two sales responses, irrespective of the evaluation criteria used. The predictive analysis reveals additional gains of 5.4% to 7% in forecast accuracy from our proposed approach, compared to the best-performing forecasting estimator that neglects model uncertainty. Despite being outperformed by the averaging LSSVR, other tree-based algorithms and SVR-type methods in our exercise generally showcase superior performance over the least squares (LS) estimator, the averaging predictive model averaging (PMA) estimator, and various penalization methods. Furthermore, the exercise confirms the benefits of optimally chosen combination weights. For example, the simple averaging LSSVR with Gaussian kernel (denoted as $\text{LSSVR}_{\text{G}}^{\text{SA}}$) is outperformed by Mallows averaging LSSVR and even some machine learning algorithms without averaging, such as the LSSVR with Gaussian kernel or random forecast.

Second, we leverage a machine learning-based tool known as the variable importance score to evaluate the significance of three promotional strategies in predicting future sales. Our findings reveal a noticeable contrast between the significance rankings from our hybrid estimator and those from the linear panel regression. The frequencies of gift and combo promotions prove more crucial under our proposed estimator, whereas the panel regression highlights the advertised discount rate and the number of combo promotions as more influential. An in-depth analysis of predictor significance, categorized by low and high previous week's store sales, suggests that the variation in rankings could be attributed to store heterogeneity. For stores with lower previous sales, promotions like advertised discounts and gifts play a more vital role in sales predictions, and store-specific dummies also emerge as important predictors. In contrast, for stores with higher previous sales, gift and combo promotions rank behind past sales records in terms of predictive

importance, and store-specific dummies are almost absent from the top 10 predictor list.

Finally, to provide guidance on the sales policy adjustment, we examine the marginal effects of those promotion predictors on the next week sales responses. This analysis is conducted via averaging LSSVR and incorporates the partial dependence (PD) plots proposed by Friedman (2001). In the case of using three promotion strategies separately, our findings reveal that combo promotions yield the largest marginal effect compared to gift and discount promotions, reaching peaks of 487.39 for the number of effective customers and 219.76 (in thousands of RMB) for sales revenues. These peak effects are observed at implementation frequencies of 240 and 210 respectively. Our finding also suggests that the ideal frequencies of gift promotions are 230 for the number of effective customers and 170 for sales revenues. In comparison to the actual statistics, our analysis indicates that there is a need to increase the number of combo promotions and decrease the number of gift promotions at the store level. Further analysis on the joint implementation of any two promotion strategies also highlights that store managers should be aware of cannibalization effects from using multiple promotion vehicles simultaneously. For instance, our study reveals that when gift and combo promotions are employed together, the most effective number of gift promotions decreases to 50 for customer visits and 100 for sales revenues, and the most effective number of combo promotions is 250 for both sales responses.

Therefore, our work is also related to the evolving literature on the modeling of promotion effects on sales. Typically, with econometric or choice models, the effects of sales promotions have been studied extensively in marketing or economics, where the main efforts are oriented toward the decomposed and dynamic effects of price promotions on grocery goods sales.[4] In contrast, our analysis employs statistical learning approaches to examine sales responses to both price and non-price promotions for a specific fashion retailer.

Our analysis also adds to the stream of works on evaluating or optimizing promotions to support retailers' strategic planning. One such study conducted by Mulhern and Leone (1991) strives to assess the profitability of discounts, taking into account the effects of substitution and complementarity between product categories. Much of the literature on optimizing promotions deals

[4]For instance, Foekens et al. (1998) use store-level scanner data to construct dynamic econometric models on sales through relating price-promotion parameters to timing and depth of past price discounts. In the same vein, Van Heerde et al. (2004) propose store-level regression models to decompose the sales promotion bump into cross-brand effects, cross-period effects and category-expansion effects. Price discount experiments are conducted by Anderson and Simester (2004) to gather household data on durable goods purchases and the opposite long-run effects for new and established customers are further revealed. While there has been limited research on non-price promotions, notable exceptions include the study by Heilman et al. (2002) on the surprise coupon. Blattberg and Scott (1994) provides a comprehensive review on sales promotions for interested readers.

with the challenges of price promotion. For example, Ferreira et al. (2016) present an efficient algorithm designed to solve a multi-product price optimization problem for an online fashion retailer.[5] In their study, Baardman et al. (2019) introduce a model that addresses the scheduling of promotion vehicles as a nonlinear bipartite matching-type problem, with the objective of maximizing profits. To our knowledge, there is a paucity of research comprehensively evaluating an assortment of promotions from a sales forecasting perspective, particularly within the realm of fashion retailing.

This paper is organized as follows. Section 2 briefly reviews existing methods, including conventional forecasting methods based on statistical models, forecast combinations applied to conventional models, and some well-known machine learning approaches. Section 3 introduces our new averaging machine learning framework. We acknowledge that the proposed framework is subject to a specific condition and discuss how widely-adopted machine learning approaches satisfy such condition in Section 3.3. We establish the asymptotic optimality of the proposed method in Section 4. Section 5 describes the data. Section 6 presents the value of our approach in sales forecasting and provides promotional insights for researchers and managers. Section 7 concludes. The Appendix provides additional details on the theoretical proofs, a more detailed review of competitor forecasting estimators, simulation results, and supplementary empirical results.

2 Empirical Techniques for Forecasting

In the context of sales forecasting, let x_{it} be one of p primary predictors (or explanatory variables) for $i = 1, ..., p$ and $t = 1, ..., T$, and y_t be the response variable of interest at period t.[6] The objective is to predict the future values of y_t with the information from a random sample of $\{y_t, X_t\}_{t=1}^{T}$, where $X_t = [1, x_{1t}, ..., x_{pt}]^\top$. Before introducing our proposed framework, we first review some existent forecasting approaches.

[5] In a related study, Caro and Gallien (2012) explore a pricing optimization issue involving clearance items for Zara, a well-known Spanish apparel retailer. Utilizing data from fast-moving consumer goods, Cohen et al. (2017) address a price promotion optimization problem.

[6] Forecasting sales has always been a major topic in the marketing and operations management literature, where regression-style models are commonly applied. For example, the influential work by Cooper et al. (1999) used the log of total units sold for each store as the dependent variable, while a variety of 67 promotion-event variables were included as predictors in the forecasting regression. Cohen et al. (2022) showed how to use data aggregation and clustering to improve retailer's demand prediction.

2.1 Linear Regressions

Linear regressions are used conventionally in the related literature as the benchmark.[7] The modeling process starts by considering a linear parametric form for the data generation process (DGP) of y_t as

$$y_t = X_t^\top \beta + e_t. \tag{1}$$

The error term is given by e_t. Assuming the values of X_{t+h} is known at time t, the h-period-ahead forecast of y_{T+h}, denoted by $\hat{y}_{T+h}$, can be expressed as

$$\hat{y}_{T+h} = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i x_{i,T+h} = X_{T+h}^\top \hat{\beta}, \tag{2}$$

where $\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p]^\top$ is an estimate of β. If β is estimated by the unrestricted least squares (LS) estimator, denoted as $\hat{\beta}^{\text{LS}}$, it can be expressed as $\hat{\beta}^{\text{LS}} = (X^\top X)^{-1} X^\top y$, where $X = [X_1, ..., X_T]^\top$ and $y = [y_1, ..., y_T]^\top$.

In the case of sales forecasting, the number of available predictors p is usually large and a significant subset of predictors is not that valuable in predicting the response variable.[8] Consequently, the out-of-sample performance of LS can be sometimes unsatisfactory. One possible solution is the penalized regression, which is capable of selecting effective predictors in order to improve forecast accuracy. The widely-applied penalized estimators include the ridge regression, the least absolute shrinkage selective operator (LASSO) and the elastic net. Further details of the above three estimators are described in Appendix B.1.

Alternatively, model averaging techniques can be used for enhancing predictive accuracy, where practitioners may have a group of plausible models and choose to combat model uncertainty by obtaining a weighted average of forecasts. Formally, assume that analysts approximate the DGP in Equation (1) with a sequence of M_T candidate models, which are defined by $y = X_{(m)}\beta_{(m)} + e_{(m)} = \mu_{(m)} + e_{(m)}$ for $m = 1, ..., M_T$ and $\mu_{(m)} = X_{(m)}\beta_{(m)}$. The selected predictors in the mth candidate model form the $T \times p_{(m)}$ matrix $X_{(m)}$, which is a subset of X with

[7] The linear multinomial logit model is adopted in Cui and Curry (2005) as the reference method for predicting consumer choices. To conduct sales forecasts in clothing industry, prior work such as Thomassey (2010) and Wong and Guo (2010) employ well-known linear models represented by Holt Winters model (Winters, 1960), ARIMA and AR(p) models.

[8] Utilizing such high dimensional data for periodic predictive analyses can be inefficient or sometimes infeasible for business practices. In an analysis of retail store-level sales forecasting, Ma et al. (2016) pointed out that the number of candidate explanatory variables can approach tens of thousands if accounting for both intra- and inter-category promotion interactions. The problem of dimensionality must be dealt with methods like penalized estimators.

$p_{(m)} \leq (p+1)$. Define the variable $\boldsymbol{w} = [w_{(1)}, ..., w_{(M_T)}]^\top$ as a weight vector in the unit simplex in $\mathbb{R}^{M_T}$,

$$\mathcal{W} \equiv \left\{ w \in [0,1]^{M_T} : \sum_{m=1}^{M_T} w_{(m)} = 1 \right\}, \tag{3}$$

where $w_{(m)}$ is the m^{th} component of w.

For model averaging, the key issue is how to assign the weights to each candidate model. Numerous optimization procedures have been developed by econometricians to tackle this.[9] Among them, the prediction model averaging (PMA) estimator is shown to perform impressively in out-of-sample experiments (Lehrer and Xie, 2017, Lehrer and Xie, 2022). Thus, it is chosen as our reference model averaging technique in the sequel. The PMA method estimates w by

$$\hat{w} = \underset{w \in \mathcal{W}}{\arg\min} \ \|y - P(w)y\|^2 + 2\hat{\sigma}^2(w)p(w),$$

where $P(w) \equiv \sum_{m=1}^{M_T} w_{(m)} P_{(m)}$ with $P_{(m)}$ being the projection matrix of $X_{(m)}$, $p(w) \equiv \sum_{m=1}^{M_T} w_{(m)} p_{(m)}$ is the effective number of parameters, and $\hat{\sigma}^2(w) = \|y - P(w)y\|^2 / (n - p(w))$ is the averaged variance. The averaged forecast of y_{T+h} from PMA is given by $X_{T+h}^\top \hat{\beta}(\hat{w})$, where

$$\hat{\beta}(\hat{w}) = \sum_{m=1}^{M_T} \hat{w}_{(m)} \Gamma_{(m)} \hat{\beta}_{(m)},$$

with $\Gamma_{(m)} = (X^\top X)^{-1} X^\top X_{(m)}$ being a $(p+1) \times p_{(m)}$ binary matrix. Matrix $\Gamma_{(m)}$ functions as a transformation matrix that expands a $p_{(m)} \times 1$ vector $\hat{\beta}_{(m)}$ to a size of $p \times 1$ by inserting zeroes.

2.2 Tree-type Machine Learning

In the literature of predicting product demands, there is ample evidence of implementing tree-type learning methods.[10] Therefore, we also include tree-type learning methods as the second

[9] Many studies have investigated questions like the choice of candidate models and weights. For example, Bates and Granger (1969) suggested to select the weights to be inversely related to estimated forecast error variances. Buckland et al. (1997) advocated choosing the weights using the Akaike Information Criteria (AIC) of all the competing models. Somewhat surprisingly, an empirically highly successful strategy is the simple averaging scheme, which assigns an equal weight to each candidate model; see Rapach et al. (2010), Elliott et al. (2013), and Claeskens et al. (2016). Non-equal weights estimated by the least squares model averaging has also become a popular choice in practice. The pioneering work is the Mallows model averaging (MMA) of Hansen (2007). Other model averaging methods include but are not limited to the jackknife model averaging (JMA) of Hansen and Racine (2012), the prediction model averaging (PMA) of Xie (2015), and the heteroskedasticity prediction model averaging (HPMA) of Zhao et al. (2016), among others. Feng et al. (2020) demonstrated that least square model averaging can also be regarded as a penalized LS regression.

[10] For the apparel and fashion industry investigated in our empirical exercise, prior works have also utilized tree-type methods mostly as demand forecasting tools of sales items with similar features. For instance, Ferreira et al. (2016)

batch of forecasting approaches. Without a linear restriction for y_t, decision trees presume

$$y_t = f(X_t) + e_t, \tag{4}$$

where the function $f(\cdot)$ can be nonlinear or even nonparametric.

There exist many algorithms to build decision trees. The building block is regression tree (RT) proposed by Breiman et al. (1984). Starting from the original data (the root node), all possible binary splits of the values for each predictor are considered and the "best split" is determined by a chosen criterion, for example, the reduction in the sum of squared residuals (SSR). Such a partitioning process can be conducted iteratively until it reaches a predetermined boundary. Many modeling parameters need to be decided or calculated *ex ante*.[11] Data in the terminal nodes (also called tree leaves) are considered to be homogeneous, hence a simple average of all the observations y_l at final leaf l is used as the fitted value. To make predictions based on X_{T+h}, we simply drop X_{T+h} down the tree and obtain the corresponding fitted value at terminal node l as the forecast $\hat{y}_{T+h}$.

As pointed out by Hastie et al. (2009), results from individual regression trees could be shaped by idiosyncratic features of the data. This drawback could be alleviated by ensemble methods that combine estimates from multiple models or trees. For example, bootstrap aggregating regression trees (also known as bagging or BAG) in Breiman (1996) first generates B bootstrap samples $\{y_t^{(b)}, X_t^{(b)}\}_{t=1}^{T}$ for $b = 1, ..., B$ from the original data, where the value of B must be predetermined. Then trees are built on each bootstrap sample and relevant forecasts $\hat{y}_{T+h}^{(b)}$ are obtained based on X_{T+h}. The variance of BAG forecasts can be large owing to the high correlation among trees. Such an issue can be circumvented by random forest (RF) of Breiman (2001). RF differs from bagging only in the set of predictors being evaluated in each tree. Random forests only take a random subset of q predictors (without replacement and $q < p$) for each splitting procedure within each tree. With both strategies, the final forecast is the simple average of forecasts from all the constructed trees. In addition to bagging and random forest, we also include a least squares boosting (LSB) algorithm (Hastie et al., 2009) in our exercise for comparison purposes. Appendix B.2 provides

report that regression trees with bagging consistently outperform the other regression models for predicting sales of first-exposure fashion styles. With real data from French textile distributor, Thomassey and Fiordaliso (2006) provide another example of using decisions trees to associate each future item with a prototype based on known descriptive criteria.

[11] These so-called tuning parameters or hyperparameters, usually contain a splitting criterion function, a minimum number of samples at a leaf node, stopping rules, etc.

additional details on the above tree-type learning strategies.

2.3 Support Vector Regression

Applications of support vector regression (SVR) and its affiliated algorithms also show high degrees of forecast accuracy without constructing tree structures.[12] Formally, Drucker et al. (1996) proposed that the SVR framework approximates $f(X_t)$ in terms of a set of basis functions $\{h_s(\cdot)\}_{s=1}^{S}$:

$$y_t = f(X_t) + e_t \approx \sum_{s=1}^{S} \beta_s h_s(X_t) + e_t, \tag{5}$$

where $h_s(\cdot)$ is implicit and can be infinite-dimensional.[13] Following Hastie et al. (2009, Chapter 12), the intercept is ignored for simplicity. The coefficients $\beta = [\beta_1, \cdots, \beta_S]^\top$ are estimated through minimization of

$$L(\beta) = \sum_{t=1}^{T} V_\varepsilon \left(y_t - f(X_t)\right) + \lambda \sum_{s=1}^{S} \beta_s^2, \tag{6}$$

where the loss function is

$$V_\varepsilon(r) = \begin{cases} 0 & \text{if } |r| < \varepsilon, \\ |r| - \varepsilon & \text{otherwise.} \end{cases}$$

The loss function V_ε is called an ε-insensitive error measure that ignores errors of size less than ε. As part of the loss function V_ε, the parameter ε is usually predefined. On the other hand, λ is a more traditional regularization parameter that can be selected by cross-validation.

Suykens and Vandewalle (1999) modified SVR which leads to solving a set of linear equations under a squared loss function. The above method, known as LSSVR, considers a similar minimization problem

$$L(\beta) = \sum_{t=1}^{T} \left(y_t - f(X_t)\right)^2 + \lambda \sum_{s=1}^{S} \beta_s^2, \tag{7}$$

where a squared loss function replaces $V_\varepsilon(\cdot)$.

[12] SVR has been implemented successfully in many business applications, particularly for those with few input variables and observations. Cui and Curry (2005) are among the early ones to analyze pros and cons of using support vector machine (SVM) for a variety of predictive tasks in marketing. SVR has also demonstrated high accuracy of forecasting grocery sales during periods with promotions, e.g. (Liu et al., 2007; Ali et al., 2009; Di Pillo et al., 2016). The recent paper of Kharfan et al. (2021) report SVM as the best performing approach for classification in a forecasting experiment for newly launched seasonal products from a fashion retail company.

[13] Note that in practice, we do not need to know β_s and the implicit function $h_s(\cdot)$. The estimation process only involves the kernel function $K(x, X_t) \equiv \sum_{s=1}^{S} h_s(x) h_s(X_t)$. See Appendix B.3 for a more comprehensive description of the estimation procedure.

We construct Lagrangian equations for (6) and (7) and solve for optimal solutions. The estimation functions for SVR and LSSVR take the following forms

$$\text{SVR} \quad : \quad \hat{f}(x) = \sum_{t=1}^{T} (\hat{\alpha}_t^* - \hat{\alpha}_t') K(x, X_t), \tag{8}$$

$$\text{LSSVR} \quad : \quad \hat{f}(x) = \sum_{t=1}^{T} \hat{\alpha}_t K(x, X_t), \tag{9}$$

for any given vector of inputs x. $\{\hat{\alpha}_t^*\}_{t=1}^T$ and $\{\hat{\alpha}_t'\}_{t=1}^T$ are the estimated Lagrangian multipliers for SVR,[14] $\{\hat{\alpha}_t\}_{t=1}^T$ are the estimated Lagrangian multipliers for LSSVR, and $K(x, X_t) \equiv \sum_{s=1}^S h_s(x) h_s(X_t)$ is a kernel function for any input vectors x and X_t. See Appendix B.3 for an extensive description of SVR-related estimation procedures.

As Equations (8) and (9) imply, no explicit forms of the basis functions are demanded during the estimation procedure. It is the kind of kernel functions that plays a crucial role in the estimation process. In this paper, we focus on the following two kernels[15]

$$\begin{aligned} \text{Linear} \quad &: \quad K(x, X_t) = x^\top X_t, \\ \text{Gaussian} \quad &: \quad K(x, X_t) = \exp\left(-\frac{\|x - X_t\|^2}{2\sigma_x^2}\right), \end{aligned}$$

where σ_x^2 is a hyperparameter. To conduct a thorough examination, we associate SVR and LSSVR with each of the above two kernels. The associated estimators are denoted as SVR_L, SVR_G, LSSVR_L, and LSSVR_G, respectively. Note that SVR_L and LSSVR_L follow a linear formulation as in (1), and the corresponding basis function is explicit.

3 Averaging Machine Learning

Applications of statistical learning are gaining popularity in the literature on retail sales forecasting. As commented by Wong and Guo (2010), the fashion sales industry is characterized by

[14] Note that additional Lagrangian multipliers are possibly required for SVR estimation, since the absolute values in $V_\varepsilon(\cdot)$ can be reformulated into two linear expressions.

[15] Another commonly used kernel is the polynomial kernel that takes the following form:

$$\text{Polynomial} : K(x, X_t) = (\gamma + x^\top X_t)^d.$$

However, results with the polynomial kernel are outperformed by the other two kernels both in our simulation experiment and in empirical exercises. Therefore, the findings are omitted for brevity and are available upon request.

uncertain customer demands and numerous driving factors. It is always desirable to have flexible forecasting methods that can handle the issue of model uncertainty. Unfortunately, most of the machine learning approaches in Section 2 do not account for this specification uncertainty.[16] Inspired by the forecast combination literature, we propose a weighted forecast combination procedure that can combine forecasts generated by machine learning methods. Note that the group of forecasts are based on various candidate models constructed from the full permutation of all potential predictors. In this way, our method can largely mitigate the problem of model uncertainty. The proposed method can be also regarded as an ensemble learning algorithm with model weights estimated by Mallows-type criteria.

3.1 Simple Averaging Machine Learning

Since equal weighted forecasting is ubiquitous in the forecast combination literature, we decide to include it as one of our reference methods. In concrete terms, an equal weight is assigned to each forecast produced by a candidate model estimated with some machine learning algorithm.[17] We denote this combination approach as simple averaging machine learning (SAML), which is universal enough to fit most machine learning algorithms.

Suppose we have a set of M_T forecasts, each of which sources from a candidate model with its own selected predictors. Denote $\hat{y}_{T+h}(m)$ as the h-step-ahead forecast of y_{T+h} based on the m^{th} model. A simple average of forecasts $\hat{y}^{\text{SA}}_{T+h}$ is given by

$$\hat{y}^{\text{SA}}_{T+h} = \frac{1}{M_T} \sum_{m=1}^{M_T} \hat{y}_{T+h}(m), \tag{10}$$

where the superscript "SA" is the abbreviation for simple averaging.[18] Clearly, the sound performance of this approach hinges on proper forecast accuracy of each candidate model. However, if

[16] Note that the RF and BAG methods naturally generate various model specifications during the sample drawing and tree-growing procedure. However, the variable selection during the splitting process is purely random. Sometimes this could yield some unexpected and confusing candidate model specifications. In the machine learning literature, ensemble learning has been put forward to boost forecast accuracy through combining outcomes from multiple algorithms. An important aspect of ensemble learning is the complete and valid quantification of model uncertainty (Liu et al., 2019).

[17] Similar to the setup in the subsequent framework, the set of candidate models is constructed from the full permutation of all potential predictors.

[18] For instance, if we focus on LSSVR with a Gaussian kernel, a single prediction $\hat{y}_{T+h}(m)$ can be computed from a candidate model estimated by LSSVR with the considered predictors. The full permutation of all predictors creates a group of potential model specifications, which later lead to a series of forecasts under LSSVR. Finally a simple average of all the above forecasts can be used as the weighted forecast of the response variable in this case.

a subset of candidate models generates fairly poor forecasts, simple averaging may fail to deliver satisfactory out-of-sample results. This postulation is sustained by the following simulation and our empirical findings.

3.2 Mallows-type Averaging Machine Learning

In an insightful paper by Ullah and Wang (2013), they argue that in a nonparametric setting one can still apply Mallows criterion to obtain the frequentist model averaging weights. The ensuing predictions correspond to each observation of the response variable by a mapping matrix.[19] In this paper, we extend their findings and propose condition C.1, under which the optimal weights can be obtained for aggregating forecasts of candidate models by a particular machine learning algorithm. We further prove that the asymptotic optimal weights are generated through minimizing the Mallows-type criterion. This new method is denoted as Mallows-type averaging machine learning (MAML).

Condition C.1 *Given the formulation of y_t as in (4), the prediction based on any input vector x must satisfy the following form*

$$\hat{f}(x) = P(x, X)y, \tag{11}$$

where y is the vector of the response variable y_t and X is the matrix of all predictors. The form of $P(x, X)$ is explicit.

Condition C.1 genuinely requires that the prediction based on any input vector x are a weighted average of obervations in y with the weights depending on both x and X in a possibly nonlinear manner.[20] Therefore, this condition does not necessarily apply to all the machine learning algorithms. However, it can be shown that there are many econometric methods fulfilling this condition.[21] An illustration of this finding with LSSVR is provided in Section 3.3. More importantly,

[19] Interested readers may refer to pages 166-168 of their paper for the concrete form of this particular matrix and the related discussion.

[20] Note that the input vector x can be a subvector of the matrix X that acts as the training set here. In this case, the prediction $\hat{f}(x)$ is actually an in-sample estimate. In other cases, x may not be retrieved from X, which results in an out-of-sample prediction.

[21] In a recent working paper by Ding et al. (2022), they provide a comprehensive demonstration on the explicit forms of the mapping matrix as in Equation (11) for many prevalent econometric approaches. Their list includes the OLS estimator, model selection, least squares model averaging, penalized methods, ridge regression, LASSO regression and LSSVR.

with the functional form in (11), we can further derive a proof of asymptotic optimality of the Mallows averaging LSSVR in Section 4.

Let us assume that the m^{th} candidate model implies the following relationship between y_t and $X_t^{(m)}$

$$y_t = f(X_t^{(m)}) + e_t^{(m)},$$

where the $p^{(m)} \times 1$ vector $X_t^{(m)}$ is a subset of X_t that includes the selected predictors, and the superscript (m) indicates predictors associated with the m^{th} submodel. Let $X_{(m)}$ be the matrix of $X_t^{(m)}$ for all t. We define $\hat{y}(m) = \hat{f}_{(m)} = P_{(m)}y$ as the prediction of y for all t by the m^{th} candidate model, where $P_{(m)}$ is a $T \times T$ matrix with the t^{th} row being $P(X_t^{(m)}, X_{(m)})$ for all $m = 1, ..., M_T$. Suppose the weight vector w satisfy $w \in \mathcal{W}$, where the set $\mathcal{W}$ is defined in (3). The weighted average prediction $\hat{f}(w)$ can be written as

$$\hat{f}(w) = \sum_{m=1}^{M_T} w_{(m)} \hat{f}_{(m)} = P(w)y,$$

where $P(w) = \sum_{m=1}^{M_T} w_{(m)} P_{(m)}$ and $w_{(m)}$ is the weight for the m^{th} candidate model.

Inspired by Hansen (2007), we propose to estimate the weight vector w by minimizing either of the following Mallows-type criteria, under the restriction of $w \in \mathcal{W}$ and various assumptions on the error term variance:

$$C_1(w) = \|y - P(w)y\|^2 + 2\sigma^2 \sum_{t=1}^{T} P_{tt}(w), \tag{12}$$

$$C_2(w) = \|y - P(w)y\|^2 + 2\sum_{t=1}^{T} \sigma_t^2 P_{tt}(w), \tag{13}$$

where $P_{tt}(w)$ is the t^{th} diagonal term in $P(w)$, σ^2 is the true error term variance under homoskedasticity, and σ_t^2 is the t^{th} true error term variance under heteroskedasticity.

Since criteria (12) and (13) include infeasible terms σ^2 and σ_t^2, we consider the alternative feasible criteria $C_1'(w)$ and $C_2'(w)$ in practice:

$$C_1'(w) = \|y - P(w)y\|^2 + 2\hat{\sigma}^2(w) \sum_{t=1}^{T} P_{tt}(w), \tag{14}$$

$$C_2'(w) = \|y - P(w)y\|^2 + 2\sum_{t=1}^{T} \hat{e}_t(w)^2 P_{tt}(w). \tag{15}$$

Criterion C_1' substitutes σ^2 with the estimated variance of averaged residuals:

$$\hat{\sigma}^2(w) = \|y - P(w)y\|^2/T, \tag{16}$$

whereas C_2' acknowledges heteroskedasticity by replacing σ_t^2 with $\hat{e}_t(w)^2$, the square of each element in the averaged residual vector $\hat{e}(w)$. The averaged residual vector $\hat{e}(w)$ is defined by

$$\hat{e}(w) = \sum_{m=1}^{M_T} w_{(m)}\hat{e}^{(m)} = (I - P(w))y, \tag{17}$$

where $\hat{e}^{(m)}$ is the residual vector for the m^{th} candidate model.

Estimating w by minimizing C_1' or C_2' is a convex optimization process. Once $\hat{w}$ is obtained with observations t for $t = 1, ..., T$ in the training set, the combined h-period ahead forecast of y_{T+h} is given by

$$\hat{y}_{T+h}^{\text{MA}} = \sum_{m=1}^{M_T} \hat{w}_{(m)}\hat{y}_{T+h}(m), \tag{18}$$

where the superscript "MA" stands for Mallows-type averaging and $\hat{y}_{T+h}(m)$ is the prediction of y at period $T+h$ by the m^{th} candidate model. Note that the optimal model weights generated in this way are mostly unequal, unlike the even weights assigned by simple averaging.

3.3 An Illustration of Mallows Averaging LSSVR

Due to the wide adoption of LSSVR in business practices,[22] we now move on to illustrate how LSSVR by Suykens and Vandewalle (1999) can be embedded into the setting of Mallows-type averaging machine learning. This is first verified technically by showing that any prediction from LSSVR obeys Condition C.1.[23]

Formally, suppose H is a $T \times r$ implicit basis matrix where $r > T$ and $H = h(X)$. The coefficients β can be estimated by minimizing the following penalized LS criterion

$$C(\beta) = \|y - H\beta\|^2 + \lambda\|\beta\|^2.$$

[22] See for instance, Yao et al. (2015) and Nazemi et al. (2018) use three variations of LSSVR to predict corporate bond recovery rates. They document significant outperformance of LSSVR compared to traditional linear regressions.

[23] As argued by Lehrer and Xie (2022), the classical SVR is incompatible with the least squares model averaging framework, because it solves an ϵ-intensive loss function.

The solution, $\hat{\beta}$, should satisfy the condition $-H^\top(y - H\hat{\beta}) + \lambda\hat{\beta} = 0$ and the in-sample prediction $\hat{f}(X)$ is therefore given by

$$\hat{f}(X) = H\hat{\beta} = \left(HH^\top + \lambda I_T\right)^{-1} HH^\top y \equiv P^{\text{LSSVR}}(X)y, \tag{19}$$

where I_T is a $T \times T$ identity matrix and

$$P^{\text{LSSVR}}(X) \equiv \left(HH^\top + \lambda I_T\right)^{-1} HH^\top \tag{20}$$

is a $T \times T$ matrix. Note that the $T \times T$ matrix $HH^\top$ is the kernel matrix with elements being $K(X_t, X_{t'}) \equiv \sum_{s=1}^{S} h_s(X_t)h_s(X_{t'})$ for different t and t'. Equation (19) proves that although the basis matrix is implicit, we can still obtain predictions complying with (11) in Condition C.1. This is due to the fact that the kernel matrix is explicit. Note that the above derivation is based on the no-intercept assumption following Hastie et al. (2009, Chapter 12). If an intercept is indispensable in model specifications, Equation (19) still holds but with a more complicated form of $P^{\text{LSSVR}}(X)$. A similar demonstration of this is provided in Appendix C.

The Mallows averaging LSSVR can then be conducted based on candidate models m for $m = 1, ..., M_T$. The initial step is to construct the $T \times T$ matrix $P_{(m)}^{\text{LSSVR}}$ for each candidate model with the chosen kernel and selected predictors.[24] Although there is no restriction on the adopted kernel function, we decide to concentrate on the Gaussian kernel for Mallows averaging LSSVR in the empirical exercise, due to its sound performance. After collecting the set of $\{P_{(m)}^{\text{LSSVR}}\}_{m=1}^{M_T}$, we can compute the averaging projection matrix $P(w)$ and plug it into Mallows criterion $C_1'(w)$ or $C_2'(w)$. The optimal w can then be estimated through a convex optimization under $C_1'(w)$ and $C_2'(w)$, respectively. Finally, the combined forecasts can be obtained based on Equation (18). For notational convenience, Mallows averaging LSSVR with Gaussian kernels under homoskedasticity and heteroskedasticity are denoted as $\text{LSSVR}_{\text{G1}}^{\text{MA}}$ and $\text{LSSVR}_{\text{G2}}^{\text{MA}}$, respectively. The estimation algorithm for Mallows averaging LSSVR is delineated in Box 1 to explain the whole procedure.

[24] With a linear kernel, the LSSVR method actually follows a linear formulation, which is equivalent to the Ridge regression discussed in Appendix B.1.

Box 1. Algorithm for Mallows Averaging LSSVR

1. For each candidate model $m = 1, ..., M_T$ with selected predictors $X^{(m)}$,

   (a) obtain the $T \times T$ kernel matrix $H_{(m)}H_{(m)}^{\top}$ with elements like

   $$K(X_t^{(m)}, X_{t'}^{(m)}) \equiv \sum_{s=1}^{S} h_s(X_t^{(m)}) h_s(X_{t'}^{(m)})$$

   for various t and t';[a]

   (b) estimate the projection matrix $P^{\text{LSSVR}}(X^{(m)})$ as in Equation (20).

2. Construct the averaging projection matrix function with the unknown vector w

   $$P(w) = \sum_{m=1}^{M_T} w_{(m)} P^{\text{LSSVR}}(X^{(m)})$$

   and plug it into criterion $C_1'(w)$ defined in (14) or criterion $C_2'(w)$ defined in (15).

   (a) For $C_1'(w)$: construct the $\hat{\sigma}^2(w)$ term following Equation (16).

   (b) For $C_2'(w)$: construct the $\hat{e}(w)$ term following Equation (17).

3. Estimate w by minimizing $C_1'(w)$ or $C_2'(w)$ under the constraint $w \in \mathcal{W}$, where $\mathcal{W}$ is defined in Equation (3).

   (a) Estimating w is a standard convex optimization process.[b]

4. Once $\hat{w}$ is obtained, the h-period-ahead forecast by Mallows averaging LSSVR is

   $$\hat{y}_{T+h}^{\text{MA}} = \sum_{m=1}^{M_T} \hat{w}_{(m)} \hat{y}_{T+h}(m),$$

   where $\hat{y}_{T+h}(m)$ denotes the LSSVR forecast of y_{T+h} by the m^{th} candidate model.

[a] The tuning parameters can either be predetermined or estimated.

[b] We mainly use the generic MATLAB function `fmincon` as the optimizer in our exercises. Other convex optimizers should work equally well.

4 Asymptotic Optimality of MAML

In this section, we first prove the asymptotic optimality of MAML under both homoskedastic and heteroskedastic error terms. We then verify that LSSVR comply with the conditions required for the proof so that it can achieve asymptotic optimality under Mallows averaging. It should be borne in mind that the proposed theorems can be applied to other machine learning algorithms, as long as relevent technical conditions are fulfilled.

To clarify some notations, suppose $\mu_t = f(X_t)$, $\mu = (\mu_1, \ldots, \mu_T)$ and $\widehat{\mu}(w) = P(w)y$. The following averaged squared error risk function is defined as

$$L_T(w) \equiv \|\widehat{\mu}(w) - \mu\|^2, \tag{21}$$

which measures the sum of squared biases between the true μ and its model averaging estimate $\widehat{\mu}(w)$. Let the expected value of the risk function be $R_T(w) = \mathbb{E}\{L_T(w)\}$ and its infimum be $\xi_T = inf_{w \in \mathcal{W}} R_T(w)$. The asymptotic optimality of MAML is achieved in the sense that the estimated risk in (21) achieves the lowest possible value asymptotically. To facilitate the later proof of Theorem 1 under homoskedasticity, we initially enumerate some necessary conditions. Please note that these conditions are reliant on the assumption of T approaching infinity.

Condition C.2 *e_t is conditionally homoskedastic and $\mathbb{E}(e_t^4|X_t) \leq v < \infty$ almost surely for $t = 1, \ldots, T$, where v is a positive constant.*

Condition C.3 $\max_m trace(P_{(m)}) = O(T^{1/2})$.

Condition C.4 *There exists a positive constant c_0 such that for all $m, s \in \{1, \ldots, M_T\}$,*

$$trace(P_{(m)}P_{(m)}^\top) \geq c_0 > 0 \quad \text{and} \quad trace(P_{(m)}P_{(s)}^\top) \geq 0$$

almost surely.

Condition C.5 *There exists a positive constant c_1 such that for all $m \in \{1, \ldots, M_T\}$,*

$$\zeta_{\max}(P_{(m)}P_{(m)}^\top) \leq c_1,$$

almost surely, where $\zeta_{\max}(B)$ denotes the largest singular value of a matrix B.

Condition C.6 *There exists a positive constant c_2 such that for all $m, j \in \{1, \ldots, M_T\}$,*

$$trace(P_{(m)}^2) \leq c_2 trace(P_{(m)}^\top P_{(m)}) \tag{22}$$

and

$$trace(P_{(j)}^\top P_{(m)} P_{(j)}^\top P_{(m)}) \leq c_2 trace(P_{(m)}^\top P_{(m)}) \tag{23}$$

almost surely.

Condition C.7 $M_T^2 \xi_T^{-1} \to 0$ *almost surely.*

Among all the above conditions, Condition C.2 establishes the boundedness of the conditional moments. Condition C.7 restricts the increasing rate of the number of candidate models M_T relative to the minimum averaging risk ξ_T. Early, Zhang (2021) proposed similar conditions, which can be traced back to their precedents in Hansen (2007) and Wan et al. (2010). Conditions C.2 and C.7 in our paper are less restrictive and more interpretable than their previous counterparts.

Conditions C.3-C.6 are high-level assumptions that restrict the trace or the largest singular value of mutual products of $P_{(m)}$ for $m = 1, \ldots, M_T$. These conditions are less interpretable but are required for proving Theorem 1. As shown in Corollary 3, these high level assumptions can be replaced with more comprehensible conditions when we study the asymptotic optimality for specific machine learning methods; for example, Mallows averaging LSSVR. For now, general cases are considered to establish the asymptotic optimality of Mallows-type averaging machine learning methods in Theorems 1 and 2.

We then continue to discuss additional conditions for the proof of Theorem 2 in the case of heteroskedastic error terms.

Condition C.8 $\mathbb{E}(e_t^4|X_t) \leq v < \infty$ *almost surely for* $t = 1, \ldots, T$, *where* v *is a positive constant.*

Condition C.9 $\max_m \max_t \iota_{tt}^{(m)} = O(T^{-1/2})$ *almost surely, where* $\iota_{tt}^{(m)}$ *is the* t^{th} *diagonal element of* $P_{(m)}$.

Similar to Condition C.2, Condition C.8 establishes the boundedness of the conditional moments under heteroskedasticity, while Condition C.9 is a replacement of Condition C.3 and concentrates on the the diagonal elements of $P_{(m)}$. Comparable assumptions can be found out in many works, see Condition (A.9) of Hansen and Racine (2012) and Assumption 4 of Zhang (2021) for instance. Moreover, Li (1987) and Andrews (1991) employed the condition $\iota_{tt}^{(m)} \leq c^\star k_{\max} T^{-1}$ where $c^\star$ is a positive constant and $k_{\max}$ is the largest number of predictors in candidate models. Obviously, when $k_{\max} = O(T^{1/2})$, their condition implies our Condition C.9.

We now define the estimated weight vectors for criteria $C_1(w)$, $C_2(w)$, $C_1'(w)$, and $C_2'(w)$ as

$$\widetilde{w} = \underset{w \in \mathcal{W}}{\arg\min}\ C_1(w) \quad , \quad \widetilde{w}' = \underset{w \in \mathcal{W}}{\arg\min}\ C_1'(w),$$

$$\overline{w} = \arg\min_{w \in \mathcal{W}} C_2(w) \quad , \quad \overline{w}' = \arg\min_{w \in \mathcal{W}} C_2'(w).$$

Theorem 1 is proposed to formally state the asymptotic optimality of the MAML estimator under homoskedastic error terms, while Theorem 2 works similarly concerning heteroskedasticity.

Theorem 1 *Assume Conditions C.1-C.7 hold. Then, as* $T \to \infty$,

$$\frac{L_T(\widetilde{w})}{inf_{w \in \mathcal{W}} L_T(w)} \xrightarrow{p} 1 \tag{24}$$

and

$$\frac{L_T(\widetilde{w}')}{inf_{w \in \mathcal{W}} L_T(w)} \xrightarrow{p} 1. \tag{25}$$

Theorem 2 *Assume Conditions C.1, C.4-C.9 hold. Then, as* $T \to \infty$,

$$\frac{L_T(\overline{w})}{inf_{w \in \mathcal{W}} L_T(w)} \xrightarrow{p} 1 \tag{26}$$

and

$$\frac{L_T(\overline{w}')}{inf_{w \in \mathcal{W}} L_T(w)} \xrightarrow{p} 1. \tag{27}$$

Complete proofs of Theorems 1 and 2 are presented in Appendix A.

To formally show the asymptotic optimality of Mallows averaging LSSVR, we then verify Conditions C.2 to C.9 for LSSVR under various kernels in Corollary 3. The most complicated case is discussed at first, where the estimated weight vector is obtained under the feasible heteroskedastic criterion $\overline{w}' = \arg\min_{w \in \mathcal{W}} C_2'(w)$. The asymptotic optimality of Mallows averaging LSSVR under other scenarios is a straightforward induction from Corollary 3.

Define $H_{(m)}$ as the basis matrix for the m^{th} candidate model underlying the LSSVR framework. Based on the conclusions of Theorem 2, we present the following corollary that establishes the asymptotic optimality of the Mallows averaging LSSVR.

Corollary 3 *Given a fixed* λ, *if there exists a positive constant* v *such that*

$$\mathbb{E}(e_t^4 | X_t) \leq v < \infty, \quad t = 1, \ldots, T, \tag{28}$$

almost surely,

$$\xi_T^{-1} M_T^2 = o(1), \tag{29}$$

almost surely, and

$$\mathcal{E}(T^{-1} H_{(m)}^\top H_{(m)}) \geq c > 0, \tag{30}$$

where $\mathcal{E}(B)$ denotes the smallest eigenvalue of a matrix B, then, as $T \to \infty$,

$$\frac{L_T(\overline{w}')}{inf_{w \in \mathcal{W}} L_T(w)} \xrightarrow{p} 1. \tag{31}$$

The Mallows averaging LSSVR is therefore asymptotically optimal.

A detailed proof of Corollary 3 is provided in Appendix A.3. Note that Conditions (28) and (29) are identical to Conditions C.8 and C.7, respectively. Condition (30) assumes that the basis matrix $H_{(m)}$ has a reasonably good behavior. A similar condition can be found in the model selection literature, for example, see Zou and Zhang (2009). They imposed the following condition on the covariate matrix of $X_{(m)}$ such that

$$\mathcal{E}(T^{-1} X_{(m)}^\top X_{(m)}) \geq c > 0. \tag{32}$$

In contrast to (32), Condition (30) is more general and makes (32) as a special case. For the LSSVR estimation with a linear kernel, Condition (30) becomes equivalent to (32) so that we can directly employ the covariate matrix of predictors rather than the basis matrix in the calculation process.

To further evaluate how the proposed framework performs, a Monte Carlo simulation experiment is carried out in Appendix D. The findings reassure that our proposed approach shows a dominant prediction performance relative to many competitive methods.

5 Data Description

We collected the weekly sales observations from stores across China of a famous footwear brand between July 5, 2021 (the 27^{th} week of the year 2021) and March 25, 2022 (the 12^{th} week of the

year 2022).[25] The data source from 35 stores across China with location details in Appendix E. Our data is an unbalanced panel consisting of 35 stores and 38 weeks. After data cleaning, our sample contains 1,168 observations. We mainly focus on two response variables at the store level: the weekly number of effective customers who make in-store purchases and the weekly sales revenue. The number of effective customers is a direct indicator reflecting the store-level traffic and the sales revenue is a crucial gauge of a store's profitability.

In our analysis, we also examine the predictors that affect future sales, which are characterized by previous sales records, one-week-ahead district-level weather forecasts and promotion activities.[26] Summary statistics are presented in Table 1. Panel A consists of three historical sales-related predictors: the number of effective customers, the units of products sold and the sales revenues from the previous week. A holiday dummy variable is also included that specifies if the upcoming week coincides with any important holiday.[27] On a weekly basis, an average store in our sample earns about 32 thousand RMB and attracts 72 customers. Panel B summarizes the one-week-ahead weather forecasts averaged over time and across stores.[28] The weather elements we use include the forecasted minimum (Temp_{Min}), maximum (Temp_{Max}) and average temperatures (Temp_{Avg}), the forecasted average (PRCP_{Avg}) and maximum precipitations (PRCP_{Max}).

The key question to be addressed here is how various promotion strategies contribute to store sales. With assistance of the brand Headquarter in China, we roughly categorize the weekly promotion activities into the three main types of promotion strategies:

(i) Promotion gift: Free gifts can be handed out by the stores to customers with purchases of certain items.

(ii) Promotion combo: Stores can offer product bundles that customers can purchase at a specific discount price compared to individual purchases.

[25] No major COVID-19 related quarantine policies were issued during this period for the cities where the stores are located.

[26] Together under the constraint of available data, we collect the predictors following several salient works in the sales forecasting literature (see for example, Cooper et al., 1999, Pauwels et al., 2002, Beheshti-Kashi et al., 2015 and Ferreira et al., 2016).

[27] In this paper, important holidays in China refer to the Mid-Autumn Festival (from September 13, 2021 to September 19, 2021), National Day (from September 27, 2021 to October 3, 2021), New Year (from December 27, 2021 to January 2, 2022), and the Spring Festival (from January 31, 2022 to February 6, 2022).

[28] The recent literature has revealed that weather significantly impacts sales of apparel and sporting goods. For example, after analyzing the data from a large European apparel retailer, de Albéniz and Belkaid (2021) find that rain and temperature have differential effects on foot traffic and successful sales of seasonal items. With the U.S. data, Roth Tran (2022) also uncovers that weather has significant persistent effects on sales, which may increase sales volatility.

(iii) Promotion discount: Reductions to various degrees on the total payment amount can be granted to purchases on a tiered fashion.

Strictly speaking, these strategies are available to all the stores at all time. However, after consulting with store managers, we find out that it is usually up to shop assistants to advise the customers on which promotion strategies to use, as the customers may not be fully aware of all the options. In addition, one purchase can involve the concurrent use of multiple promotions if applicable. For example, one may purchase product combos and items with free gifts at the same time.

To further illustrate the details, a gift item is usually an accessory or a tag-along that values much less than the main item. Promotion combo always involves buying a bundle of items together, where the promotion item is not an accessory and is restrictive to be of the same type as the purchased item (possibly with varying colors or sizes). Both gift and combo promotions are constrained to certain items decided by the store managers, which leaves the customers with limited choices. Promotion discount on the other hand is more flexible. Although the discount percentages are tiered by the total amount of purchase, there are usually no constraints on the applicable items. Please refer to Appendix F for an in-depth exploration of the data source and the strategies employed for promotions.

We counted the weekly implementation of each promotion strategy at each store and constructed three corresponding promotion predictors P_{Gift}, P_{Combo}, and $\mathrm{P}_{Discount}$. Since percentages of promotion discount have direct impacts on store revenues, we also compute the promotion "off-rate" (in decimals) for each store by comparing the calculated revenue under original prices with the actual revenue under discounted prices:

$$\text{Off-rate} = 1 - \frac{\text{Revenue after Promotion}}{\text{Revenue before Promotion}} \in [0, 1).$$

The off-rate therefore measures the revenue loss induced by discounts. Common sense tells us that a higher off-rate is supposed to attract more customers.[29] To further capture the influence from consumer expectations, we incorporate advertised discount rates (Ad. Discount) measured

[29]The marketing literature documents that the long-run effects of promotion on sales may be more complex, which usually work through mechanisms such as forward buying, selection, customer learning and increased deal sensitivity. Using data on durable goods, Anderson and Simester (2004) find that deeper price discounts boost future purchases from first-time customers, while they work in the opposite way on established customers. The findings based on packaged goods reveal that consumers become more price and promotion sensitive over time because of frequent promotions (Mela et al., 1997).

in decimals as the last promotion predictor, which provides store-by-store information about the upcoming discounts for the next week.

As implied in Panel C of Table 1, the discount promotion ($P_{Discount}$) is the most frequently offered promotion averaging 26.32 times per week. The gift promotion (P_{Gift}) is the next in the ranking with an average of 14.80 times per week. The combo strategy is the least applied promotion with a mean of 8.13 times each week. The minimum values for three promotions are all zeros, implying that some store at certain period did not initiate any promotion. The mean off-rate is 14.31% with a maximum of 73.27%. In contrast, the mean (26.08%) and the median (30%) of the advertised discount rates are much higher.

Table 1: Summary Statistics

Predictor	Mean	Median	Maximum	Minimum	Std. Dev.	Skewness	Kurtosis
Panel A: Previous Sales							
lag(Customer)	71.9366	42.0000	891.0000	1.0000	91.0936	3.7673	23.5356
lag(Unit) (in thousands)	0.3471	0.1955	4.0200	0.0000	0.4515	3.4833	19.5301
lag(Revenue) (in thousands of RMB)	31.6695	16.7515	388.8820	0.0000	44.1079	3.7405	21.9786
Holiday	0.1858	0.0000	1.0000	0.0000	0.3891	1.6158	3.6107
Panel B: Weather Forecast							
$Temp_{Min}$	45.3100	41.8000	79.9000	-10.3000	18.7927	0.0242	2.2282
$Temp_{Max}$	72.7527	69.8000	105.8000	28.2000	16.7334	0.0271	1.9510
$Temp_{Avg}$	58.0601	54.6500	89.2000	11.4000	17.0308	0.0722	2.0922
$PRCP_{Avg}$	12.4449	0.0814	99.9900	0.0000	24.6330	1.9418	5.5802
$PRCP_{Max}$	25.7124	0.4050	99.9900	0.0000	43.3014	1.1327	2.2845
Panel C: Promotion Activities							
P_{Gift}	14.7997	2.0000	567.0000	0.0000	50.6769	6.6348	54.3578
P_{Combo}	8.1318	5.0000	73.0000	0.0000	9.7679	2.8853	13.5832
$P_{Discount}$	26.3185	12.0000	603.0000	0.0000	46.1139	5.6446	51.7779
Off-rate	0.1431	0.1422	0.7327	0.0000	0.1175	0.6543	3.6295
Ad. Discount	0.2608	0.3000	0.6580	0.0000	0.1466	-0.1372	3.4371

Notes. This table reports summary statistics of all the predictors for store-level sales. The temperature (Temp) is measured in the Fahrenheit scale and the cap of precipitation (PRCP) is set to a level of 99.99mm. Lag(Customer), lag(Unit) and lag(Revenue) respectively denote the number of effective customers, the units of products sold and the sales revenues from the previous week. Among them, revenues are measured in thousands of RMB and units are measured in thousands.

6 Empirical Exercise

In this section, we conduct an empirical analysis using store-level observations described in Section 5. We first run a conventional panel regression on the number of effective customers and sales revenues. Then the empirical evidence demonstrating nonlinearity between the response variables and the predictors is presented. In the next step, we conduct an out-of-sample comparison using 13 estimators commonly encountered in the literature and two averaging algorithms proposed in Section 3. Finally, we rank the predictors to understand their relative importance and

quantify the marginal effects of various promotions on the two sales responses with the proposed algorithms.

6.1 In-Sample Estimation

We conduct a linear panel regression on the full sample with an explicit control of store-level and holiday fixed effects. The estimation results are reported in Table 2. As it can be seen, the historical sales predictors in Panel A are all significantly positive for explaining effective customer visits and sales revenues next week, with the only exception of negative impacts from lagged sales units. The results here also reveal positive serial correlations of the two response variables. Our results in Panel B support the literature that weather prominently affects the customer behavior and sales of apparel products.[30] Extremely high temperature and heavy rain both cause significant declines in effective customers, whereas warmer average temperature encourages customers to visit and make purchases. As for sales revenues, average temperature has a clearly positive impact, of about \$642.7 per degree. The impact of rain (PRCP_{Max}) is marginally negative (-\$49.2 per mm). Generally speaking, effective customers and sales are mainly driven by temperature, so that more warmer days increase the sales probability of footwear products.

The marginal effects of three promotion strategies are displayed in Panel C of Table 2.[31] The number of gift promotion (P_{Gift}) and the off-rate are insignificant for explaining customer visits and store revenues next week. The number of combo promotion (P_{Combo}) and advertised discount rates (Ad.Discount) are significantly positive, with combo promotion manifesting larger impacts on both customer visits and store revenues. In contrast, promotion by discount ($\text{P}_{Discount}$) is significantly negative at 1% level.

The signs of three promotion strategies are in agreement with the related literature. Among them, $\text{P}_{Discount}$ is the most straightforward type of price promotion, with which customers are tempted to make more purchases in order to obtain higher discount rates. Such kind of promotions

[30] de Albéniz and Belkaid (2021) conduct a study with daily observations of casual apparel sales from 13 European markets. Similar to ours, they also find that average temperature matters the most for sales conversions and rain plays a negative role for customer visits. In the same vein, Roth Tran (2022) also conclude that extreme heat events lead to significant declines in sales of a U.S. apparel and sporting good brand.

[31] To verify the causality between promotion predictors and sales responses, we conduct a quasi-experimental analysis in Appendix G. The signs of the estimated treatment effects on the two response variables are compatible with the results of the fixed effect regression shown in Table 2. This sustains the subsequent marginal effect analysis of three promotion predictors.

Table 2: Estimates of Panel Regression

Predictor	# of Customers	Sales Revenue (in thousands)
Panel A: Previous Sales		
lag(Customer)	0.9693***	0.2222***
	(0.1015)	(0.0571)
lag(Unit)	-103.7215***	-41.9062***
	(20.2131)	(12.8212)
lag(Revenue)	0.5827**	0.6819***
	(0.2789)	(0.2107)
Holiday	12.7230***	5.4595***
	(3.2034)	(1.9937)
Panel B:Weather Forecast		
Temp_{Min}	-0.4654	-0.2254
	(0.3238)	(0.1577)
Temp_{Max}	-0.6349***	-0.1209
	(0.2270)	(0.1188)
Temp_{Avg}	1.7400***	0.6427**
	(0.5676)	(0.2852)
PRCP_{Avg}	0.0965	0.0441
	(0.0726)	(0.0379)
PRCP_{Max}	-0.1091**	-0.0492*
	(0.0454)	(0.0252)
Panel C: Promotion Activities		
P_{Gift}	0.0514	0.0396
	(0.0777)	(0.0460)
P_{Combo}	0.6526***	0.4836***
	(0.2382)	(0.1754)
$\text{P}_{Discount}$	-0.3114***	-0.1724***
	(0.0573)	(0.0331)
Off Rate	-36.2748	-16.9878
	(23.8213)	(13.6894)
Ad. Discount	39.9540***	18.2193***
	(11.2342)	(5.1909)
Panel D: Goodness-of-Fit		
$\bar{R}^2$	0.7311	0.7250
Adj. R^2	0.7195	0.7132

Notes. This table reports coefficient estimates and relevant statistics for a panel regression with fixed effects on the sales data. Numbers in parentheses represent the heteroskedasticity-robust standard errors. Superscripts *, **, *** indicate that the associated coefficients are significant at levels of 10%, 5%, and 1%, respectively. The centered R^2 and the adjusted R^2 values are further reported in the last two rows.

can lead to the so called 'post-promotion' dips on next week's sales and store traffic.[32] On the other hand, P_{Combo} has the potential to prompt the customers to make another purchase in close proximity to the one primed by the combo.[33] The gift promotion in our case is relatively ineffective to boost the store sales for the next week. This result is anticipated, since free gifts only strengthen

[32]Other papers in the marketing literature use data of household goods, and have reached the same conclusion that the post-promotion effect on sales is usually negative, through the channels of selection (Neslin and Shoemaker, 1989) or forward buy (Krishna 1992, 1994). Blattberg and Scott (1994) also find out that forward buying is more prevalent for durable goods.

[33]This can occur due to two possible scenarios: either not all the items in the combo are desired by the customers or some products in the combo stimulate purchases for items cognitively related to the one primed by the combo. Interested readers can refer to Heilman et al. (2002) for a study on the surprise coupon, which functions in a similar manner as the promotion combo.

the current purchase intention of main products (Gilbert and Jackaria, 2002).

6.2 Nonlinearity

The results from the panel regressions in Section 6.1 provide mean estimates across all the stores. The linearity assumption behind may be too restrictive to describe the actual relationship between the predictors and the response variables. Moreover, these estimates ignore the obvious heterogeneity across the stores. To motivate the application of machine learning and our proposed approaches, it is helpful to show that the assumption of a linear and additive relationship between y_t and $X_t = [x_{1t}, ..., x_{pt}]^\top$ is too strong.[34]

To do so, it begins with considering a fully nonparametric function that relates y_t to X_t. However, if the hypothetical relationship was imposed, one would face the curse of dimensionality because of the overwhelming 48 predictors. Therefore, we instead consider a partially linear model

$$y_t = Z_{1t}^\top \beta + g(Z_{2t}) + e_t, \tag{33}$$

where Z_{1t} is a $k \times 1$ vector, β is the associated $k \times 1$ coefficient vector, Z_{2t} is a $q \times 1$ vector (i.e., $q = p - k$), $g(\cdot)$ is an infeasible, possibly nonlinear function, and e_t is the error term.

To make the dimensionality manageable, a small q shall be used. Following Li and Racine (2007), an infeasible estimator of β by LS is described by

$$\tilde{\beta} = \left(\sum_{t=1}^{T} \tilde{Z}_{1t} \tilde{Z}_{1t}^T \right)^{-1} \sum_{t=1}^{T} \tilde{Z}_{1t} \tilde{y}_t, \tag{34}$$

where $\tilde{Z}_{1t} = Z_{1t} - \mathbb{E}(Z_{1t}|Z_{2t})$ and $\tilde{y}_t = y_t - \mathbb{E}(y_t|Z_{2t})$.

In practice, the conditional expectations in (34) can be consistently estimated using the kernel method:

$$\begin{aligned} \hat{y}_t &\equiv \hat{\mathbb{E}}(y_t|Z_{2t}) = T^{-1} \sum_{j=1}^{T} y_t K_h(Z_{2t}, Z_{2j}) \Big/ \hat{f}(Z_{2t}), \\ \hat{Z}_{1t} &\equiv \hat{\mathbb{E}}(Z_{1t}|Z_{2t}) = T^{-1} \sum_{j=1}^{T} Z_{1j} K_h(Z_{2t}, Z_{2j}) \Big/ \hat{f}(Z_{2t}), \end{aligned}$$

[34] This is in the spirit of work by Gutierrez et al. (2008) which clearly show that neural network models generally perform better than the traditional time series methods at forecasting lumpy demands from a Mexican electronics distributor.

where $\hat{f}(Z_{2t}) = T^{-1}\sum_{j=1}^{T} K_h(Z_{2t}, Z_{2j})$, $K_h(Z_{2t}, Z_{2j}) = \prod_{s=1}^{q} h_s^{-1} k\left(\frac{Z_{2ts} - Z_{2js}}{h_s}\right)$ with $k(\cdot)$ being the kernel function and h_s being the bandwidth for the s^{th} element in Z_{2t}.

The presence of the random denominator $\hat{f}(Z_{2t})$ can cause some technical difficulties when deriving the asymptotic distribution of the feasible estimator β. We consider a simple approach that trims out observations of which the denominator is small and such a feasible estimator of β is defined by

$$\hat{\beta} = \left(\sum_{t=1}^{T}(Z_{1t} - \hat{Z}_{1t})(Z_{1t} - \hat{Z}_{1t})^{\top}\right)^{-1} \sum_{t=1}^{T}(Z_{1t} - \hat{Z}_{1t})(y_t - \hat{y}_t)\mathbb{I}_t\left(\hat{f}(Z_{2t}) \geq b\right), \tag{35}$$

where $\mathbb{I}_t(\cdot)$ is an indicator function that equals one if the input argument is true and zero otherwise. The trimming parameter $b = b_n > 0$ and satisfies $b_n \to 0$ asymptotically. Once $\hat{\beta}$ is obtained and the condition $Z_{2t} = z$ holds, the nonparametric components can be estimated consistently by

$$\hat{g}(z) = \frac{\sum_{j=1}^{T}(y_t - Z_{1t}^{\top}\hat{\beta})K_h(z, Z_{2t})}{\sum_{j=1}^{T} K_h(z, Z_{2t})}. \tag{36}$$

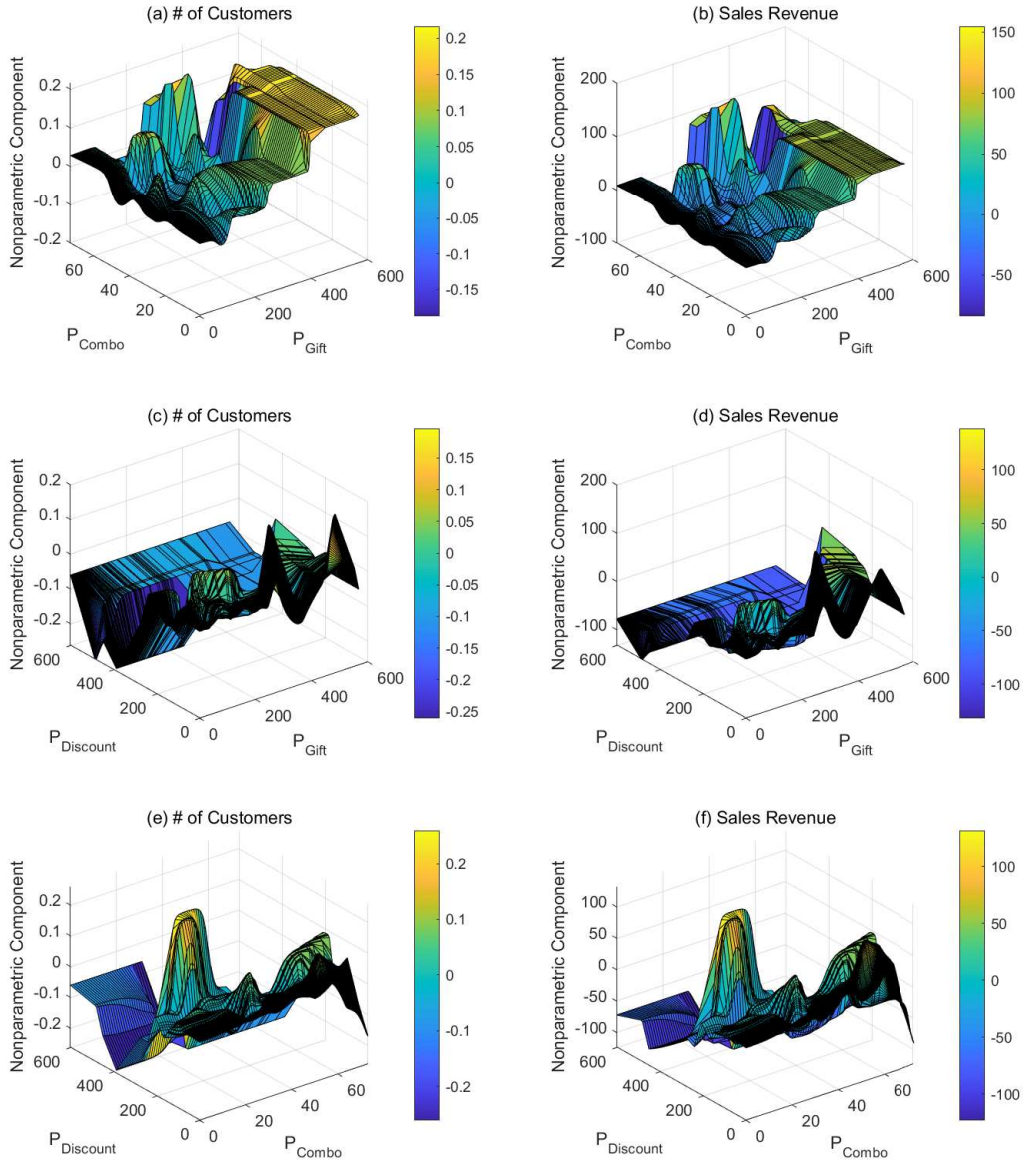
We next concentrate on showing the nonparametric relationship between the response variables and the three promotion strategies. The other variables are deemed linear as in Equation (33). The Gaussian kernel is chosen with the optimal bandwidth $\hat{h}_z = 1.06\hat{\sigma}_z T^{-1/(q+4)}$ for each component of Z_{2t}. We consider a full combination of any two promotion strategies (i.e. $q = 2$) as the predictors for $g(Z_{2t})$ so as to generate a range of surface plots for each response variable.

The estimated $g(Z_{2t})$ is plotted in Figure 1. The two columns of the subplots correspond to each response variable indicated in the subtitles. Each subplot suggests that the relationship between the sales response variable and the selected promotion predictors is obviously nonlinear. The above exercise confirms that it is inadequate to use a linear model, at least in our sample. This finding calls for the adoption of nonlinear estimators such as machine learning algorithms or our proposed estimators. For comparison, we still include several conventional estimators with linear formulations in the empirical exercise.

6.3 Forecasting Exercise

In this section, we consider predicting the number of effective customers and sales revenues with the predictors defined in Section 5. Table 3 outlines 15 forecasting estimators used in the exercise.

Figure 1: Evidence Illustrating Nonlinearity between Sales Response and Promotion Strategies



Notes. This figure presents the surface plots for nonparametric components of the partially linear model in (33). Panels (a), (c) and (e) describe the relationship between the number of effective customers and a full combination of any two promotion variables; Panels (b), (d) and (f) provide similar plots for the sales revenues.

Table 3: List of Estimators Evaluated in the Forecasting Exercise

Method Abbreviation	Detailed Description
Panel A: Linear Estimators	
LS	Unrestricted ordinary least squares estimator.
LASSO	The least absolute shrinkage selective operator by Tibshirani (1996).
RIDGE	The ridge estimator.
EN	The elastic net by Zou and Hastie (2005).
PMA	The prediction model averaging method by Xie (2015).
Panel B: Tree-type Algorithms	
RT	Regression tree by Breiman et al. (1984).
LSB	Least squares regression boosting of Hastie et al. (2009, Chapter 10).
BAG	Bootstrap aggregation tree by Breiman (1996).
RF	Random forest by Breiman (2001).
Panel C: SVR-type Methods	
SVR_L	Support vector regression of Drucker et al. (1996) with linear kernel.
SVR_G	Support vector regression of Drucker et al. (1996) with Gaussian kernel.
$LSSVR_G$	Least squares support vector regression of Suykens and Vandewalle (1999) with Gaussian kernel.
Panel D: Averaging LSSVR	
$LSSVR_G^{SA}$	Simple averaging LSSVR with Gaussian kernel discussed in Section 3.
$LSSVR_{G1}^{MA}$	Mallows-type averaging LSSVR with Gaussian kernel under homoskedasticity.
$LSSVR_{G2}^{MA}$	Mallows-type averaging LSSVR with Gaussian kernel under heteroskedasticity.

Notes. The abbreviation for each estimator is listed in the first column of each panel of Table 3. The second column of each panel provides a brief description of the estimator and more details are presented in Appendix B.

We conduct a rolling window forecasting exercise with window length (WL) of 20 weeks. We use the data from all the stores as the training set to train our estimators and predict the sales for each store one-week-ahead.[35] The performance of most learning techniques depend on tuning parameters. In practice, these tuning parameters can be either predetermined (see for example, our simulation experiment in Appendix D) or estimated. The latter approach usually relies on re-sampling techniques like cross-validation (CV), which is far more computationally intensive than the former. Bergmeir et al. (2018) argued that the standard five-fold CV is valid in autoregressive models with uncorrelated errors. Unless otherwise indicated, results in the sequel are produced based on tuning parameters set by five-fold CV via grid search. We have also tried three-fold CV and ten-fold CV, and the findings are qualitatively intact. Principal tuning parameters and their ranges for grid search (if there are any) are listed below:

1. The regularization parameter $\lambda \in \{0, 0.01, 0.02, ..., 100\}$ for LASSO, RIDGE, EN, SVR-type, and averaging LSSVR ;

2. The mixing parameter $\alpha \in \{0.1, 0.5, 0.9\}$ for EN;

[35] Conducting an individual store-level prediction analysis is impractical in our case due to the limited number of observations available per store, with a maximum of only 38 observations.

3. The minimum leaf size spans from 1 to 10 for all the tree-type algorithms;

4. The number of learning cycles is set to $B = 100$ for all the ensemble methods;

5. The number of selected predictors for RF is picked from $\{\lfloor p/4 \rfloor, \lfloor p/3 \rfloor, \lfloor p/2 \rfloor\}$;

6. The hyperparameter $\sigma_x^2 \in \{0.1, 0.2, ..., 100\}$ for the Gaussian kernel.

Note that both BAG and RF do not require five-fold CV to select optimal values for their tuning parameters. Since BAG and RF rely on the bootstrap resampling process, we use only a fraction of total observations each time we draw a bootstrap sample. The rest of the observations are called the out-of-bag (OOB) observations, which act as an ideal test set to evaluate the constructed tree and choose optimal tuning parameters. See Appendix B.2 for a detailed discussion.

Another important issue for the proposed averaging LSSVR is the number of potential models in the candidate model set. The conventional approach to construct the model set is to employ a full combination of all the p predictors, which yields a large number of 2^{48} candidate models in our case. To reduce the computational burden, we designate the predictors of previous sales and the store-wise dummy variables for each candidate model. The predictors of previous sales are central to explaining the two responses and the store-wise dummy variables are vital for describing the store-wise heterogeneous effects. The above step reduces the total number of candidate models considerably from 2^{48} to 2^9 (512).

Inspired by Yuan and Yang (2005), we further perform the below model screening process on the 512 candidate models. Each candidate model is evaluated by the following criterion:

$$C(s) = \|y - P_{(s)}y\|^2 + 2\hat{\sigma}^2_{(s)} \sum_{t=1}^{T} P^{(s)}_{tt},$$

for $s = 1, ..., 512$, where $P_{(s)}$ stands for the projection matrix $P(x, X)$ of the s^{th} candidate model, $\hat{\sigma}^2_{(s)}$ is the variance of estimated error terms, and $P^{(s)}_{tt}$ represents the t^{th} diagonal term in $P_{(s)}$. The candidate models are then ranked according to the values of $C(s)$ in ascending order and the top M models are selected. The value of M is set to be $M = 12$ in our exercise.[36]

The results on prediction comparison for the two response variables are reported in Table 4.[37]

[36] We have also tried alternative values of M larger than 12 and found that the results are almost identical but with longer computational time.

[37] Note that we additionally compare the out-of-sample accuracy of the pooling LS regression with that of individual store-wise LS regressions in Appendix H.1. The forecast accuracy of the pooling regression is much higher in all cases.

Our proposed Mallows averaging LSSVR ($\text{LSSVR}_{\text{G1}}^{\text{MA}}$ or $\text{LSSVR}_{\text{G2}}^{\text{MA}}$) yields the best performance whether evaluated by SDFE or MAFE for both response variables. The $\text{LSSVR}_{\text{G1}}^{\text{MA}}$ seems to outperform $\text{LSSVR}_{\text{G2}}^{\text{MA}}$ by a small margin in three out of the four cases. The LSSVR_{G} approach and RF offer the second best performance if assessed by SDFE, where LSSVR_{G} performs almost as well as RF. Accounting for model uncertainty leads to gains of 5.4% to 7% between LSSVR_{G} and the best performing Mallows averaging LSSVR. It is also interesting to see that the naive averaging $\text{LSSVR}_{\text{G}}^{\text{SA}}$ is dominated by some machine learning algorithms without averaging, for example LSSVR_{G} or RF. The above finding confirms that even in the applications with machine learning algorithms, asymptotically optimal model weights are still valuable for improving prediction accuracy.

Table 4: Prediction Comparison for Sales Response Variables

Method	SDFE		MAFE	
	# of customers	Sales Revenue	# of customers	Sales Revenue
Panel A: Linear Estimators				
LS	33.3456	15.9741	19.2915	9.1839
LASSO	33.4121	16.2313	17.5141	8.4573
Ridge	56.2600	25.2084	40.6904	17.2078
EN	32.0415	16.2797	17.3296	8.4586
PMA	33.1692	15.6923	18.9601	8.9257
Panel B: Tree-type Algorithms				
RT	33.3738	14.5412	18.8833	8.3124
LSB	34.3734	15.4653	22.2613	9.4349
BAG	30.7845	14.4747	16.8696	7.7437
RF	29.8302	14.1460	16.6494	7.6539
Panel C: SVR-type Methods				
SVR_{L}	31.4006	14.9748	16.6136	7.6861
SVR_{G}	43.5639	20.3372	29.6473	13.7660
LSSVR_{G}	29.3678	14.2863	17.3527	8.1551
Panel D: Averaging LSSVR				
$\text{LSSVR}_{\text{G}}^{\text{SA}}$	29.5084	14.4194	16.8030	7.7404
$\text{LSSVR}_{\text{G1}}^{\text{MA}}$	**27.1913**	**13.5177**	**16.4139**	7.6265
$\text{LSSVR}_{\text{G2}}^{\text{MA}}$	27.2540	13.5205	16.5322	**7.6230**

Notes. The results of the prediction exercise for the two sales responses are reported in this table. A full description of each estimator in the first column is provided in Table 3. The risks of the forecasting exercise are evaluated by SDFE and MAFE presented in the left and right panels, respectively. Bold numbers denote the estimator with the lowest risk and thus the best performance in that column of the table.

Although outperformed by averaging LSSVR, other tree-based algorithms and SVR-type methods in our exercise mostly manifest dominating performance over the LS estimator, the averaging PMA estimator and penalization methods, with the sole exception of LSB and SVR_{G}. The approaches such as bagging and RF bear the feature of ensemble learning and noticeably exceed other tree-based algorithms and conventional support vector regressions. Taking these findings together, we conclude that there are obvious improvements from using averaging machine learn-

ing approaches that can simultaneously accommodate nonlinearity and model specification uncertainty.

6.4 Relative Importance of Predictors

The managerial practice calls for an interpretable understanding of how the predictors correlate with the sales response variables and a further identification of which predictors weigh more in conducting forecasts. To offer valuable insights for refining promotional policies, our analysis specifically concentrates on the predictors associated with promotions. Following the empirical strategy in Lehrer and Xie (2022), we evaluate the relative importance of a specific predictor by measuring the loss in accuracy if that particular predictor is excluded from the model. Such exclusion can be approximated by a random permutation of the predictor which aims to destroy the correlation between the predictor and the response variable.

Taking RF as an example, we grow each tree with its respective randomly drawn bootstrap sample. The observations that are excluded from the bootstrap sample are called the Out-of-Bag sample (OOB), which becomes the perfect evaluation set since the related observations do not take part in the training process. For a given predictor, we first randomly permute it in the OOB sample that generates the modified OOB sample. Then the gap between prediction errors of the tree on the modified OOB sample and the untouched OOB sample is calculated. This process is reiterated for each tree and each predictor so that the average of these gaps in prediction errors across all OOB samples is computed. The averaged gap provides an estimate of the overall decrease in accuracy that the permutation of removing a specific predictor induces. Therefore it acts as the variable importance score for each predictor by which we can rank its relative importance. The most crucial predictors are the ones yielding the highest scores.[38] The detailed computational algorithm is delineated in Box 2. In the main exercise, we employ the regular bootstrap method and set $B = 1000$.[39]

[38] See Ishwaran (2007) for a theoretical argument of tree-based variable importance measures.

[39] We also consider the moving block bootstrap method formulated by Künsch (1989) as an alternative resampling method that draws blocks of observations in order to preserve the chronological order at the store-level in our data. Table A4 reveals that the findings by regular and block bootstrap methods are quite similar. See Appendix H for additional details.

Box 2. Algorithm for Computing Variable Importance Score

1. Take a random sample of size T with replacement from the data (bootstrapping).[a]
2. Train the forecasting strategy using the bootstrap sample and obtain the OOB sample.
3. Apply the trained strategy to the OOB sample and obtain the SDFE, denoted as SDFE^0.
4. For predictor $i = 1, ..., p$,
   (a) randomly permute the predictor i in the OOB sample;
   (b) apply the trained strategy to the modified OOB sample and compute the SDFE^i;
   (c) calculate the associated gap^i by $\text{gap}^i = \text{SDFE}^i - \text{SDFE}^0$.
5. Repeat steps 1 to 4 for B times and calculate the Score^i for each predictor i by

$$\text{Score}^i = \frac{1}{B}\sum_{b=1}^{B} \text{gap}_b^i,$$

   where gap_b^i is the estimated gap for the predictor i with the b^{th} bootstrap sample.
6. Rank the predictors by Score^i. The most important predictor yields the highest score.

[a] The alternative approach is the moving block bootstrap that resamples blocks of observations instead of individual observations.

Table 5: Top 10 Most Important Predictors by Three Forecasting Estimators

Ranking	# of Customers	Sales Revenue	# of Customers	Sales Revenue	# of Customers	Sales Revenue
	Random Forest		LSSVR_G		$\text{LSSVR}_\text{G1}^\text{MA}$	
1	lag(Customer)	lag(Revenue)	lag(Revenue)	lag(Revenue)	lag(Revenue)	lag(Revenue)
2	lag(Revenue)	lag(Customer)	lag(Unit)	lag(Unit)	lag(Customer)	lag(Unit)
3	Temp_{Avg}	Temp_{Avg}	lag(Customer)	lag(Customer)	lag(Unit)	lag(Customer)
4	Temp_{Max}	lag(Unit)	P_{Gift}	P_{Combo}	P_{Gift}	P_{Combo}
5	lag(Unit)	P_{Gift}	P_{Combo}	P_{Gift}	P_{Combo}	P_{Gift}
6	P_{Gift}	Temp_{Max}	Temp_{Avg}	Temp_{Max}	Temp_{Max}	Temp_{Max}
7	Holiday	Temp_{Min}	Temp_{Min}	Off Rate	Temp_{Avg}	Temp_{Avg}
8	Temp_{Min}	Off Rate	Temp_{Max}	Temp_{Avg}	$\text{P}_{Discount}$	Off Rate
9	Off Rate	P_{Combo}	Off Rate	Temp_{Min}	Off Rate	$\text{P}_{Discount}$
10	P_{Combo}	Holiday	$\text{P}_{Discount}$	$\text{P}_{Discount}$	Temp_{Min}	Temp_{Min}

Note. This table reports the ranking of the 10 most important predictors for the number of effective customers and sales revenues by three better-performing estimators.

As RF, LSSVR_G, and $\text{LSSVR}_\text{G1}^\text{MA}$ demonstrate promising performance in the forecasting exercise, we decide to compute variable importance scores based on each of these three estimators. The top 10 most important predictors under each estimator are listed in Table 5. We find that the top 10 most important predictors selected by RF, LSSVR_G, and $\text{LSSVR}_\text{G1}^\text{MA}$ are nearly the same, which cover three groups of predictors summarized in Table 1: previous sales, weather forecast, and promotion activities. However, different forecasting estimators produce slightly different rankings of predictor importance. All three estimators consider the previous sales variables to

be the most critical group. This aligns with the finding from the panel regression that all previous sales predictors are significantly positive.

The second group of crucial predictors differs slightly across the three estimators. With RF, we find that the weather-related predictors are crucial for forecasting the sales responses, particularly for predicting the number of effective customers. With LSSVR_{G} and $\text{LSSVR}_{\text{G1}}^{\text{MA}}$, promotion activities such as gifts and combos are a bit more valuable to predict customer visits and revenues than weather factors. There is a marked contrast between the ranking of promotion predictors by the three estimators and the outcomes from the panel regression. For the results with LSSVR_{G} and $\text{LSSVR}_{\text{G1}}^{\text{MA}}$, we observe that the frequencies of gift (P_{Gift}) and combo (P_{Combo}) promotions are among the top 5 predictors for the two responses, where P_{Gift} may play a bigger role for the effective customers and P_{Combo} matters more for explaining sales revenues. Even with RF, these two promotion predictors are among the top 10 important predictors. Advertised discount rate does not belong to the top 10 important predictors. However, if judged by coefficient magnitudes from the panel regression, combo promotions and advertised discount rates seem to have higher predictive power.

To provide further motivations for the findings above, we next examine if there is any variation in the predictor significance across the entire distribution of sales responses. The rationale behind this analysis stems from the hypothesis that promotions of varying types may yield diverse effects on stores possessing dissimilar characteristics. For each sales response, we divide the store-level data into two portions by the median of lagged sales response variables (i.e., the median of lag(Customer) for # of customers and the median of lag(Revenue) for sales revenue). In this manner, we categorize the data into two groups: those exhibiting low sales performance from the previous week and those displaying high sales performance from the same period.

Table 6 reports the ranking outcomes for low and high previous sales measures, respectively. On average, we notice that promotions such as advertised discount rates and gifts are more critical for stores with low previous sales. Moreover, we find that store-wise dummy variables are also among the top 10 principal predictors for stores with low previous sales. This discovery suggests the possibility that the distribution of sales responses may be associated with underlying store attributes, thereby emphasizing the necessity to account for store-specific heterogeneity in our analysis. On the other hand, the predictor ranking for stores with high previous sales in Panel B is quite similar to those in Table 5. Promotion predictors become less important than the previous

sales and store-wise dummy variables almost disappear from the list. In summary, the aforementioned analysis indicates that promotional strategies, such as advertised discounts and gifts, play a pivotal role in predicting sales for stores with low prior sales. However, the importance of such promotions diminishes for stores that have demonstrated high previous sales.

Table 6: Heterogeneity in the Relative Importance of Predictors

Ranking	# of Customers	Sales Revenue	# of Customers	Sales Revenue	# of Customers	Sales Revenue
	Random Forest		$LSSVR_G$		$LSSVR_{G1}^{MA}$	
Panel A: Low Sales Measure from Last Week						
1	Ad. Discount	Ad. Discount	lag(Customer)	$Temp_{Max}$	lag(Customer)	lag(Customer)
2	lag(Customer)	lag(Revenue)	P_{Gift}	lag(Customer)	P_{Gift}	lag(Revenue)
3	$Temp_{Avg}$	$Temp_{Max}$	$Temp_{Max}$	lag(Revenue)	Holiday	$Temp_{Max}$
4	$Temp_{Max}$	lag(Customer)	Ad. Discount	lag(Unit)	$Temp_{Max}$	Ad. Discount
5	Holiday	lag(Unit)	store13	store17	lag(Unit)	lag(Unit)
6	lag(Revenue)	$Temp_{Avg}$	lag(Unit)	Ad. Discount	store13	P_{Gift}
7	$Temp_{Min}$	$Temp_{Min}$	store21	store13	Ad. Discount	store13
8	lag(Unit)	Off Rate	Holiday	P_{Gift}	store21	store23
9	Off Rate	Holiday	store31	Holiday	store18	store21
10	$P_{Discount}$	$PRCP_{Avg}$	store18	store15	store31	$Temp_{Avg}$
Panel B: High Sales Measure from Last Week						
1	lag(Customer)	lag(Revenue)	lag(Revenue)	lag(Revenue)	lag(Revenue)	lag(Revenue)
2	lag(Revenue)	lag(Customer)	lag(Unit)	lag(Unit)	lag(Unit)	lag(Unit)
3	$Temp_{Avg}$	P_{Gift}	lag(Customer)	lag(Customer)	lag(Customer)	lag(Customer)
4	$Temp_{Max}$	lag(Unit)	P_{Gift}	P_{Combo}	P_{Gift}	P_{Combo}
5	P_{Gift}	$Temp_{Avg}$	$Temp_{Avg}$	$Temp_{Max}$	P_{Combo}	P_{Gift}
6	Holiday	$Temp_{Max}$	P_{Combo}	P_{Gift}	$Temp_{Max}$	$Temp_{Max}$
7	lag(Unit)	$Temp_{Min}$	Off Rate	Off Rate	$Temp_{Avg}$	Off Rate
8	$Temp_{Min}$	Off Rate	$Temp_{Min}$	$Temp_{Avg}$	Off Rate	$Temp_{Avg}$
9	Off Rate	Holiday	$Temp_{Max}$	$Temp_{Min}$	$P_{Discount}$	$P_{Discount}$
10	P_{Combo}	P_{Combo}	$P_{Discount}$	$P_{Discount}$	store27	$Temp_{Min}$

Note. With RF, $LSSVR_G$ and $LSSVR_{G1}^{MA}$, this table presents the top 10 most important predictors for effective customers and sales revenues in each subsample. Results for the low and high sales measures from last week are reported in Panels A and B, respectively.

6.5 Marginal Effects of Promotion Predictors

Evaluating the incremental impacts of various promotions and adjusting sales strategies accordingly is a critical issue of interest to management boards. Meanwhile, understanding and visualizing the role of each predictor for the predicted response is of paramount importance in many supervised learning applications (Apley and Zhu, 2020). In the linear regressions, the marginal effects of promotion predictors on the two sales responses can be readily captured by coefficient estimates, as shown in Table 2. However, the coefficients are mean estimates that ignore heterogeneity and possible nonlinearity between the predictors and the response variables. In this section, using the partial dependence (PD) plots by Friedman (2001), we demonstrate the marginal effects of promotion predictors with averaging LSSVR.

The PD plot reveals the dependence between the target response and a set of input features

of interest, marginalizing over the values of all other input features. Let X_S be the set of input features of interest and X_C be its complement. The partial dependence of the response function f at a point x_S is defined as:

$$pd_{X_S}(x_S) \equiv \mathbb{E}_{X_C}[f(x_S, X_C)] = \int f(x_S, x_C) p(x_C) dx_C \approx \frac{1}{T} \sum_{i=1}^{T} f(x_S, x_C^{(i)}),$$

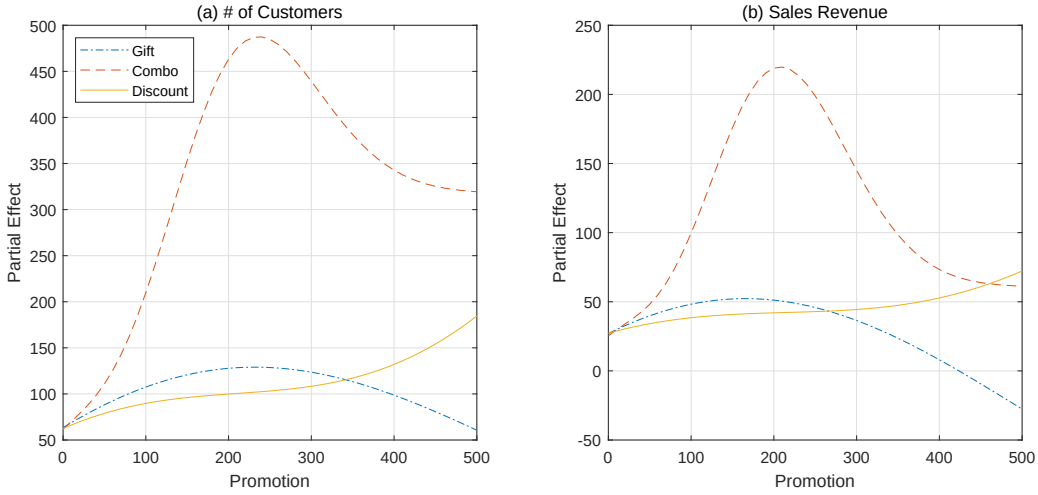
where $f(x_S, X_C)$ is the response function for a given sample, of which the values are defined by x_s and x_C for the respective features in X_S and X_C, and $x_C^{(i)}$ is the value of the i^{th} sample for the features in X_C.

For each value of x_S, calculating the PD requires a full pass over the whole dataset, which can be computationally intensive. In our exercise, we first calculate the PD of each promotion predictor on the two sales responses using $\text{LSSVR}_{\text{G}}^{\text{MA}}$ as the response function. For a specific x_S, its value is presumed to range from 0 to 500, increasing in increments of $d = 10$.[40] The PD plots of the three promotion predictors on the two sales responses are presented in Figure 2. The two subplots correspond to the two sales responses, respectively. In each subplot, the horizontal and vertical axes represent the value ranges of promotion predictors and the estimated partial effects, respectively. The marginal effects of P_{Gift}, P_{Combo}, and $\text{P}_{Discount}$ are indicated by the dash-dotted line, the dashed line, and the solid line, respectively.

The marginal effects of three promotion predictors by $\text{LSSVR}_{\text{G1}}^{\text{MA}}$ display notable dynamics across the spectrum of promotional frequencies. Several interesting findings are worth stressing. First, the marginal effects of combo promotions consistently surpass those of the other two promotion strategies, regardless of the sales responses involved. The peak marginal effect for the number of customers reaches 487.39, while for sales revenues, it tops at 219.76. These peak effects occur at implementation frequencies of 240 and 210, respectively. Second, the marginal effects of gift promotions encompass a more limited scope and occasionally become negative (for instance, within the range of $[40, 125]$ for the number of customers and $[-30, 52]$ for sales revenues). The optimal frequencies of gift promotions are 230 for the number of customers and 170 for sales revenues. In stark contrast to the panel regression findings, the marginal effects of discount promotions consistently exhibit positive and increasing trends for both sales responses. However, their impacts are not as pronounced as gift promotions within the initial ranges of implementa-

[40] Note that the promotion predictors in our data are all non-negative integers. The range of x_S considered in this exercise is consistent with statistics of actual data.

Figure 2: PD Plots of Individual Promotion Predictors on Sales Responses by $\text{LSSVR}_{\text{G1}}^{\text{MA}}$



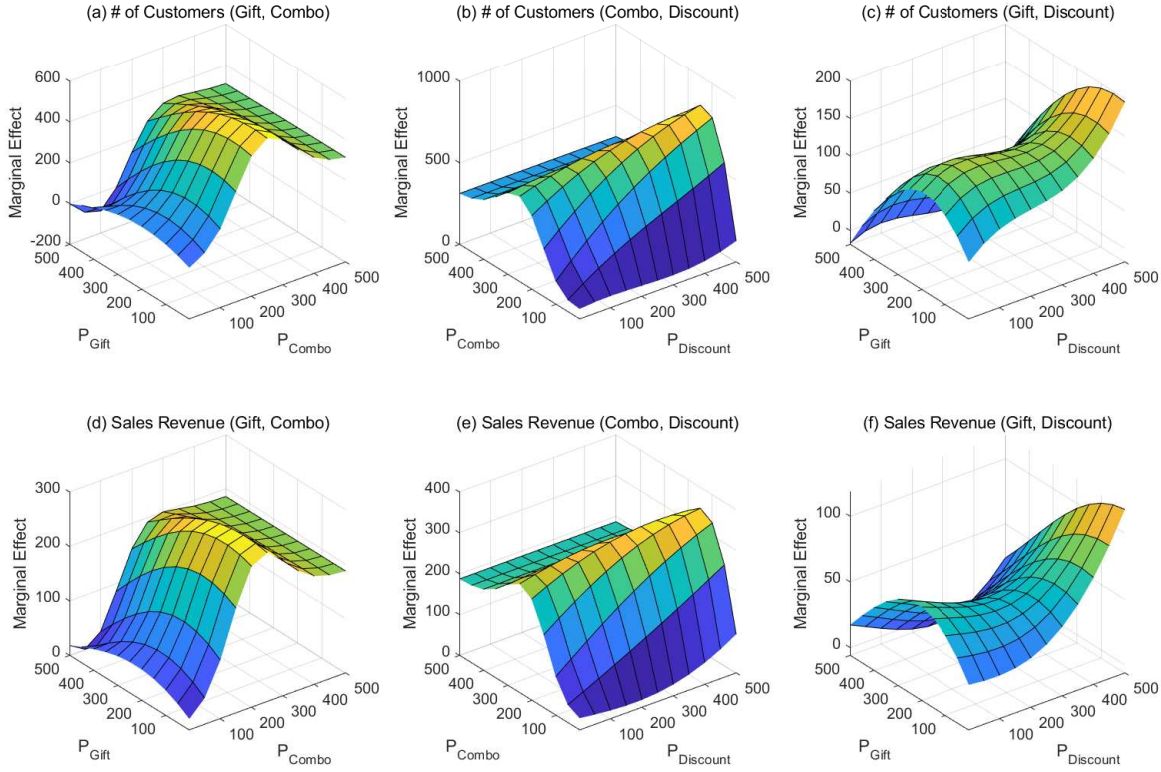
Notes. This figure plots the estimated PD between each promotion predictor (P_{Gift}, P_{Combo}, and $\text{P}_{Discount}$) and each sales response (the number of effective customers and sales revenues) by $\text{LSSVR}_{\text{G1}}^{\text{MA}}$. The horizontal and vertical axes represent the ranges of promotion predictors and the estimated partial effects, respectively.

tion frequencies (ranging from 0 to 350 times for the number of customers and from 0 to 250 times for sales revenues). Beyond these thresholds, the marginal effects of discounts exceed those of gift promotions.

In order to examine the combined effects of various promotions, we delve deeper into the interactive marginal effects of different promotional strategies. Since there are three promotion strategies, we plot the PD estimates based on the following three combinations: (i) (Gift, Combo), (ii) (Gift, Discount), and (iii) (Combo, Discount). For each pair of promotion predictors, we evaluate a range of their values within $[0, ..., 500]$ using increments of $d = 50$ to manage the computational load effectively. Figure 3 presents the surface plots based on the three pairs of promotion predictors and the response function by $\text{LSSVR}_{\text{G1}}^{\text{MA}}$. The three columns represent the plots for each pair of promotion strategies and the two rows correspond to the results for each response variable. In each subplot, the three axes represent the ranges of two promotion predictors and the estimated interactive PD, respectively.

The interactive partial dependence (PD) plots clearly illustrate that, when promotional strategies are jointly employed, the most effective implementation frequencies linked to peak marginal effects can be identified in certain instances. As the effective number of gift promotion is 50 for the number of customers and 100 for sales revenues, the effective number of combo promotions

Figure 3: Interactive PD Plots of Promotion Combinations on Sales Responses



Notes. This figure presents the estimated interactive PD between each pair of promotions ((Gift, Combo), (Combo, Discount), and (Gift, Discount)) and each of the two sales responses (the number of effective customers and the sales revenue). The estimates are calculated using the $\text{LSSVR}_{\text{G1}}^{\text{MA}}$ estimator.

is observed to be 250 for both response metrics. Deviating from this combo promotion number would reduce its impact and also counteract the benefits of gift promotions. This finding diverges somewhat from the other two promotion pairings. The remaining subplots indicate that it is consistently advantageous to elevate the implementation number of discount promotion when it is combined with either combo or gift promotions at their respective optimal frequencies.

The analysis above yields several valuable recommendations and managerial insights when associated with statistics in Table 1. First, store managers should consider designing more combo promotions, as the statistics suggest that even the maximum value of combo promotions at certain stores amounts to only 73 times, which is significantly lower than the estimated optimal frequencies for combo promotions. Second, the statistics indicate that some stores may be overusing gift promotions, as the maximum number of gift promotions in the sample is 567, considerably higher than the most effective frequencies suggested by the analysis. Lastly, our analysis underscores that store managers should be cognizant of potential cannibalization effects resulting from the concurrent use of multiple promotional vehicles.[41] For the joint implementation of gift and combo promotions, the most effective frequencies of gift promotions are decreased for both sales responses, while the most effective number for combo promotions is marginally increased for both sales responses. In contrast, when involving discounts with other promotion types, they can be applied at their individual best frequencies without impacting one another.

7 Conclusion

As documented by Liu et al. (2013), the sales of fashion products, including apparel, shoes, and beauty items, are greatly influenced by multiple factors, such as seasonality, fashion trends, weather, sales promotions, and macroeconomic conditions. Hence, predictive analytics must take into account a diverse range of explanatory variables, which inherently introduces the challenge of model uncertainty. Despite extensive research into complex and non-linear relationships between fashion sales series, the issue of model specification uncertainty has not been sufficiently addressed in previous studies.

[41] Cannibalization in marketing also refers to a reduction in sales volume, sales revenue, or market share of one product as a result of the introduction of a new product by the same producer. This effect has been studied and quantified in the literature. For more details on the measurement of this effect, please refer to Van Heerde et al. (2004) and McColl et al. (2020) for the cases of grocery sales.

To add to the relevant literature, we invent a new type of averaging forecasting estimator, which is characterized by provable optimal weighted forecasts and an array of submodel predictions generated by machine learning algorithms. Noted that the set of sub-models are constructed through the full permutation of all potential predictors. We present a rigorous proof that demonstrates the optimality of our model weights, under the condition that the predictions derived from any input vector can be mathematically expressed as a weighted average of observations in the response variables. Thereby our method demonstrates sufficient versatility to fit various commonly used machine learning techniques, offering the advantages of mitigating model uncertainty and simultaneously accommodating nonlinearity in the data.

Embedding LSSVR into our averaging estimator, we study empirically how our approach contributes to sales forecasting and promotion evaluation in a fashion retailor setting, where the sales associates' choices of promotional strategies drive the store sales. Using weekly store-level data tracking all the 35 stores in China for an internationally renowned footwear brand, we primarily examine two sales responses-the weekly number of effective customers and the weekly sales revenues, and three promotional predictors-gift, combo and discount promotions in our analysis. In our evaluation, we assess a comprehensive set of 15 competitive forecasting approaches, including linear estimators, recursive partitioning estimators, support vector regressions and the proposed averaging LSSVR.

We find that our proposed averaging LSSVR achieves superior forecast accuracy for the two sales responses, surpassing other top-performing forecasting estimators that do not account for model uncertainty. The improvement in forecast accuracy ranges from 5.4% to 7% in comparison. The exercise further confirms the value of optimal combination weights, demonstrating that averaging LSSVR and certain machine learning algorithms without averaging, such as LSSVR with Gaussian kernel or random forecast, outperform simple averaging LSSVR with Gaussian kernel.

Based on our averaging LSSVR estimator, we investigate the significance of three promotion predictors in impacting store sales for the upcoming week. Our study uncovers contrasting significance rankings between our averaging estimator and the linear panel regression, with gift and combo promotions showing greater significance in our hybrid approach while the panel regression emphasizes the importance of advertised discount rate and combo promotions. An further analysis of predictor significance based on sub-samples of low and high previous week's store sales indicates that store heterogeneity may contribute to the variation in rankings.

We conduct an additional exercise to measure the marginal effects of three promotion predictors. The results highlight that combo promotions have the most significant impact, with peak effects observed at implementation frequencies of 240 and 210 for the effective customer visits and sales revenues, respectively. Our finding also suggests that the ideal frequencies of gift promotions are 230 for the number of effective customers and 170 for sales revenues. In comparison to the actual predictor statistics, our analysis indicates that there is a need to increase the number of combo promotions and decrease the number of gift promotions at the store level. However, when considering the joint implementation of gift and combo promotions, caution is advised because of cannibalization effects. In such cases, it is advisable to decrease the number of gift promotions to 50 for customer visits and 100 for sales revenues, while slightly increasing the number of combo promotions to 250.

There are several avenues to explore for future research. Firstly, further investigations could be conducted to examine the applicability and effectiveness of the proposed averaging estimator in other retail settings and industries. This would help validate its versatility and robustness across various contexts. Additionally, a more in-depth analysis could be undertaken to understand the underlying mechanisms that drive the impact of various promotional strategies on sales. Furthermore, considering the evolving nature of the fashion industry and the emergence of new promotional approaches (such as influencer marketing and social media campaigns), future research could integrate these novel strategies into the forecasting and promotion evaluation framework. Overall, these future research directions would contribute to enhancing the accuracy and effectiveness of sales forecasting and promotion decision-making in the fashion retail industry.

References

ALI, O. G., S. SAYIN, T. VAN WOENSEL, AND J. FRANSOO (2009): "SKU Demand Forecasting in the Presence of Promotions," *Expert Systems with Applications*, 36, 12340–12348.

ANDERSON, E. T. AND D. I. SIMESTER (2004): "Long-Run Effects of Promotion Depth on New Versus Established Customers: Three Field Studies," *Marketing Science*, 23, 4–20.

ANDREWS, D. (1991): "Asymptotic Optimality of Generalized C_L, Cross-validation, and Generalized Cross-validation in Rregression with Heteroskedastic Errors," *Journal of Econometrics*, 47, 359–377.

APLEY, D. W. AND J. ZHU (2020): "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models," *Journal of the Royal Statistical Society Series B*, 82, 1059–1086.

AU, K.-F., T.-M. CHOI, AND Y. YU (2008): "Fashion Retail Forecasting by Evolutionary Neural Networks," *International Journal of Production Economics*, 114, 615–630.

BAARDMAN, L., M. C. COHEN, K. PANCHAMGAM, G. PERAKIS, AND D. SEGEV (2019): "Scheduling Promotion Vehicles to Boost Profits," *Management Science*, 65, 50–70.

BARNARD, G. A. (1963): "New Methods of Quality Control," *Journal of the Royal Statistical Society. Series A (General)*, 126, 255–258.

BATES, J. M. AND C. W. J. GRANGER (1969): "The Combination of Forecasts," *Operational Research Quarterly*, 20, 451–468.

BEHESHTI-KASHI, S., H. R. KARIMI, K.-D. THOBEN, M. LÜTJEN, AND M. TEUCKE (2015): "A Survey on Retail Sales Forecasting and Prediction in Fashion Markets," *Systems Science & Control Engineering*, 3, 154–161.

BERGMEIR, C., R. J. HYNDMAN, AND B. KOO (2018): "A Note on the Validity of Cross-validation for Evaluating Autoregressive Time Series Prediction," *Computational Statistics & Data Analysis*, 120, 70–83.

BLATTBERG, R. AND N. SCOTT (1994): *Sales Promotion Concepts, Methods, and Strategies.*, Englewood Cliffs, NJ: Prentice-Hall.

BOX, G. E. P., G. M. JENKINS, G. C. REINSEL, AND G. M. LJUNG (2015): *Time Series Analysis: Forecasting and Control,*, Wiley Series in Probability and Statistics, Hoboken, NJ, USA: John Wiley & Sons.

BREIMAN, L. (1996): "Bagging Predictors," *Machine Learning*, 26, 123–140.

——— (2001): "Random Forests," *Machine Learning*, 45, 5–32.

BREIMAN, L., J. FRIEDMAN, AND C. J. STONE (1984): *Classification and Regression Trees*, Chapman and Hall/CRC.

BUCKLAND, S. T., K. P. BURNHAM, AND N. H. AUGUSTIN (1997): "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618.

CARO, F. AND J. GALLIEN (2012): "Clearance Pricing Optimization for a Fast-Fashion Retailer," *Operations Research*, 60, 1404–1422.

CLAESKENS, G., J. R. MAGNUS, A. L. VASNEV, AND W. WANG (2016): "The Forecast Combination Puzzle: A Simple Theoretical Explanation," *International Journal of Forecasting*, 32, 754 – 762.

CLEMEN, R. T. (1989): "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting*, 5, 559–583.

COHEN, M. C., N.-H. Z. LEUNG, K. PANCHAMGAM, G. PERAKIS, AND A. SMITH (2017): "The Impact of Linear Optimization on Promotion Planning," *Operations Research*, 65, 446–468.

COHEN, M. C., R. ZHANG, AND K. JIAO (2022): "Data aggregation and demand prediction," *Operations Research*, 70, 2597–2618.

COOPER, L. G., P. BARON, W. LEVY, M. SWISHER, AND P. GOGOS (1999): "PromoCast^{TM}: A New Forecasting Method for Promotion Planning," *Marketing Science*, 18, 301–316.

CUI, D. AND D. CURRY (2005): "Prediction in Marketing Using the Support Vector Machine," *Marketing Science*, 24, 595–615.

DE ALBÉNIZ, V. M. AND A. BELKAID (2021): "Here Comes the Sun: Fashion Goods Retailing under Weather Fluctuations," *European Journal of Operational Research*, 294, 820–830.

DE BRABANTER, K., J. DE BRABANTER, J. A. K. SUYKENS, AND B. DE MOOR (2011): "Approximate Confidence and Prediction Intervals for Least Squares Support Vector Regression," *IEEE Transactions on Neural Networks*, 22, 110–120.

DI PILLO, G., V. LATORRE, S. LUCIDI, AND E. PROCACCI (2016): "An Application of Support Vector Machines to Sales Forecasting under Promotions," *4OR*, 14, 309–325.

DING, W., S. F. LEHRER, AND T. XIE (2022): "Algorithms for Predictive Analytics: Communication, Privacy and Weights," *NBER Working Paper*.

DONG, X., Z. YU, W. CAO, Y. SHI, AND Q. MA (2020): "A Survey on Ensemble Learning," *Frontiers of Computer Science*, 14, 241–258.

DRUCKER, H., C. J. C. BURGES, L. KAUFMAN, A. J. SMOLA, AND V. VAPNIK (1996): "Support Vector Regression Machines," in *Advances in Neural Information Processing Systems 9*, ed. by M. C. Mozer, M. I. Jordan, and T. Petsche, MIT Press, 155–161.

ELLIOTT, G., A. GARGANO, AND A. TIMMERMANN (2013): "Complete Subset Regressions," *Journal of Econometrics*, 177, 357 – 373.

ELLIOTT, G. AND A. TIMMERMANN (2016): *Economic Forecasting*, Princeton University Press.

FENG, Q., J. G. SHANTHIKUMAR, AND M. XUE (2022): "Consumer Choice Models and Estimation: a Review and Extension," *Production and Operations Management*, 31, 847–867.

FENG, Y., Q. LIU, AND R. OKUI (2020): "On the Sparsity of Mallows Model Averaging Estimator," *Economics Letters*, 187, 108916.

FERREIRA, K. J., B. H. A. LEE, AND D. SIMCHI-LEVI (2016): "Analytics for an Online Retailer: Demand Forecasting and Price Optimization," *Manufacturing & Service Operations Management*, 18, 69–88.

FOEKENS, E. W., P. S.H. LEEFLANG, AND D. R. WITTINK (1998): "Varying Parameter Models to Accommodate Dynamic Promotion Effects," *Journal of Econometrics*, 89, 249–268.

FRANK, C., A. GARG, L. SZTANDERA, AND A. RAHEJA (2003): "Forecasting Women's Apparel Sales Using Mathematical Modeling," *International Journal of Clothing Science and Technology*, 15, 107–125.

FREUND, Y. AND R. E. SCHAPIRE (1997): "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, 55, 119 – 139.

FRIEDMAN, J. H. (2001): "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, 29, 1189 – 1232.

GAO, Y., X. ZHANG, S. WANG, T. T.-L. CHONG, AND G. ZOU (2019): "Frequentist Model Averaging for Threshold Models," *Annals of the Institute of Statistical Mathematics*, 71, 275–306.

GENRE, V., G. KENNY, A. MEYLER, AND A. TIMMERMANN (2013): "Combining Expert Forecasts: Can Anything Beat the Simple Average?" *International Journal of Forecasting*, 29, 108 – 121.

GILBERT, D. AND N. JACKARIA (2002): "The Efficacy of Sales Promotions in UK Supermarkets: A Consumer View," *International Journal of Retail & Distribution Management*, 30, 315–322.

GUTIERREZ, R. S., A. O. SOLIS, AND S. MUKHOPADHYAY (2008): "Lumpy Demand Forecasting Using Neural Networks," *International Journal of Production Economics*, 111, 409–420.

HANSEN, B. E. (2007): "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189.

HANSEN, B. E. AND J. S. RACINE (2012): "Jackknife Model Averaging," *Journal of Econometrics*, 167, 38–46.

HASTIE, T., R. TIBSHIRANI, AND J. FRIEDMAN (2009): *The Elements of Statistical Learning*, Springer Series in Statistics, New York, NY, USA: Springer New York Inc.

HEILMAN, C. M., K. NAKAMOTO, AND A. G. RAO (2002): "Pleasant Surprises: Consumer Response to Unexpected In-Store Coupons," *Journal of Marketing Research*, 39, 242–252.

HOETING, J. A., D. MADIGAN, A. E. RAFTERY, AND C. T. VOLINSKY (1999): "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–401.

ISHWARAN, H. (2007): "Variable Importance in Binary Regression Trees and Forests," *Electronic Journal of Statistics*, 1, 519–537.

JAMES, G., D. WITTEN, T. HASTIE, AND R. TIBSHIRANI (2017): *An Introduction to Statistical Learning: with Applications in R*, New York, NY, USA: Springer-Verlag New York, Inc.

KHARFAN, M., V. CHAN, AND T. FIRDOLAS EFENDIGIL (2021): "A Data-driven Forecasting Approach for Newly Launched Seasonal Products by Leveraging Machine-learning Approaches," *Annals of Operations Research*, 303, 159–174.

KREISS, J.-P. AND S. N. LAHIRI (2012): "Bootstrap Methods for Time Series," in *Time Series Analysis: Methods and Applications, Volume 30*, ed. by T. S. Rao, S. S. Rao, and C. Rao, North Holland, chap. 1, 3–26.

KRISHNA, A. (1992): "The Normative Impact of Consumer Price Expectations for Multiple Brands on Consumer Purchase Behavior," *Marketing Science*, 11, 266–286.

——— (1994): "The Impact of Dealing Patterns on Purchase Behavior," *Marketing Science*, 13, 351–373.

KÜNSCH, H. R. (1989): "The Jackknife and the Bootstrap for General Stationary Observations," *Annals of Statistics*, 17, 1217–1241.

LEHRER, S. F. AND T. XIE (2017): "Box Office Buzz: Does Social Media Data Steal the Show from Model Uncertainty When Forecasting for Hollywood?" *The Review of Economics and Statistics*, 99, 749–755.

——— (2022): "The Bigger Picture: Combining Econometrics with Analytics Improve Forecasts of Movie Success," *Management Science*, 68, 189–210.

LI, K.-C. (1987): "Asymptotic Optimality for C_p, C_L, Cross-validation and Generalized Cross-validation: Discrete Index Set," *Annals of Statistics*, 15, 958–975.

LI, Q. AND J. RACINE (2007): *Nonparametric Econometrics: Theory and Practice*, Princeton University Press, 1 ed.

LIU, J., J. PAISLEY, M.-A. KIOUMOURTZOGLOU, AND B. COULL (2019): "Accurate Uncertainty Estimation and Decomposition in Ensemble Learning," in *Advances in Neural Information Processing Systems*, ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, Curran Associates, Inc., vol. 32, 1–12.

LIU, N., S. REN, T.-M. CHOI, C.-L. HUI, AND S.-F. NG (2013): "Sales Forecasting for Fashion Retailing Service Industry: A Review," *Mathematical Problems in Engineering*, 2013, 1–9.

LIU, Y., Y. YIN, J. GAO, AND C. TAN (2007): "Demand Forecasting by Using Support Vector Machine," in *Third International Conference on Natural Computation (ICNC 2007)*, vol. 3, 272–276.

MA, S., R. FILDES, AND T. HUANG (2016): "Demand Forecasting with High Dimensional Data: The Case of SKU Retail Sales Forecasting with Intra- and Inter-category Promotional Information," *European Journal of Operational Research*, 249, 245–257.

MCCOLL, R., R. MACGILCHRIST, AND S. RAFIQ (2020): "Estimating Cannibalizing Effects of Sales Promotions: The Impact of Price Cuts and Store Type," *Journal of Retailing and Consumer Services*, 53, 101982.

MELA, C. F., S. GUPTA, AND D. R. LEHMANN (1997): "The Long-Term Impact of Promotion and Advertising on Consumer Brand Choice," *Journal of Marketing Research*, 34, 248–261.

MULHERN, F. J. AND R. P. LEONE (1991): "Implicit Price Bundling of Retail Products: A Multiproduct Approach to Maximizing Store Profitability," *Journal of Marketing*, 55, 63–76.

NAZEMI, A., K. HEIDENREICH, AND F. J. FABOZZI (2018): "Improving Corporate Bond Recovery Rate Prediction Using Multi-factor Support Vector Regressions," *European Journal of Operational Research*, 271, 664–675.

NESLIN, S. A. AND R. W. SHOEMAKER (1989): "An Alternative Explanation for Lower Repeat Rates after Promotion Purchases," *Journal of Marketing Research*, 26, 205–213.

NI, Y. AND F. FAN (2011): "A Two-stage Dynamic Sales Forecasting Model for the Fashion Retail," *Expert Systems with Applications*, 38, 1529–1536.

PAUWELS, K., D. M. HANSSENS, AND S. SIDDARTH (2002): "The Long-Term Effects of Price Promotions on Category Incidence, Brand Choice, and Purchase Quantity," *Journal of Marketing Research*, 39, 421–439.

RAO, C. R. (1973): *Linear Statistical Inference and Its Applications*, vol. 2, Wiley New York.

RAPACH, D. E., J. K. STRAUSS, AND G. ZHOU (2010): "Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy," *The Review of Financial Studies*, 23, 821–862.

REID, D. J. (1968): "Combining Three Estimates of Gross Domestic Product," *Economica*, 35, 431–444.

ROTH TRAN, B. (2022): "Sellin' in the Rain: Weather, Climate, and Retail Sales," *Working Paper*.

SAGI, O. AND L. ROKACH (2018): "Ensemble Learning: A Survey," *WIREs Data Mining and Knowledge Discovery*, 8, e1249.

SUN, Z.-L., T.-M. CHOI, K.-F. AU, AND Y. YU (2008): "Sales Forecasting Using Extreme Learning Machine with Applications in Fashion Retailing," *Decision Support Systems*, 46, 411–419.

SUYKENS, J. AND J. VANDEWALLE (1999): "Least Squares Support Vector Machine Classifiers," *Neural Processing Letters*, 9.

SUYKENS, J. A. K., T. V. GESTEL, J. D. BRABANTER, B. D. MOOR, AND J. VANDEWALLE (2002): *Least Squares Support Vector Machines*, Singapore: World Scientific Publishing Company.

THOMASSEY, S. (2010): "Sales Forecasts in Clothing Industry: The Key Success Factor of the Supply Chain Management," *International Journal of Production Economics*, 128, 470–483.

THOMASSEY, S. AND A. FIORDALISO (2006): "A Hybrid Sales Forecasting System Based on Clustering and Decision Trees," *Decision Support Systems*, 42, 408–421.

TIBSHIRANI, R. (1996): "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58, 267–288.

TIMMERMANN, A. (2006): "Chapter 4 Forecast Combinations," Elsevier, vol. 1 of *Handbook of Economic Forecasting*, 135 – 196.

ULLAH, A. AND H. WANG (2013): "Parametric and Nonparametric Frequentist Model Selection and Model Averaging," *Econometrics*, 1, 157–179.

VAN HEERDE, H. J., P. S. H. LEEFLANG, AND D. R. WITTINK (2004): "Decomposing the Sales Promotion Bump with Store Data," *Marketing Science*, 23, 317–334.

VAPNIK, V. N. (1996): *The Nature of Statistical Learning Theory*, New York, NY, USA: Springer-Verlag New York, Inc.

WAN, A. T., X. ZHANG, AND G. ZOU (2010): "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156, 277–283.

WHITTLE, P. (1960): "Bounds for the Moments of Linear and Quadratic Forms in Independent Variables," *Theory of Probability & Its Applications*, 5, 302–305.

WINTERS, P. R. (1960): "Forecasting Sales by Exponentially Weighted Moving Averages," *Management Science*, 6, 324–342.

WONG, W. AND Z. GUO (2010): "A Hybrid Intelligent Model for Medium-term Sales Forecasting in Fashion Retail Supply Chains Using Extreme Learning Machine and Harmony Search Algorithm," *International Journal of Production Economics*, 128, 614–624.

XIE, T. (2015): "Prediction Model Averaging Estimator," *Economics Letters*, 131, 5–8.

YAO, X., J. CROOK, AND G. ANDREEVA (2015): "Support Vector Regression for Loss Given Default Modelling," *European Journal of Operational Research*, 240, 528–538.

YESIL, E., M. KAYA, AND S. SIRADAG (2012): "Fuzzy Forecast Combiner Design for Fast Fashion Demand Forecasting," *2012 International Symposium on Innovations in Intelligent Systems and Applications*, 1–5.

YUAN, Z. AND Y. YANG (2005): "Combining Linear Regression Models: When and How?" *Journal of the American Statistical Association*, 100, 1202–1214.

ZHANG, X. (2010): "Model Averaging and Its Applications," *Ph.D. Thesis*, Academy of Mathematics and Systems Science, Chinese Academy of Sciences.

——— (2021): "A New Study on Asymptotic Optimality of Least Squares Model Averaging," *Econometric Theory*, 37, 388–407.

ZHAO, S., X. ZHANG, AND Y. GAO (2016): "Model Averaging with Averaging Covariance Matrix," *Economics Letters*, 145, 214 – 217.

ZOU, H. AND T. HASTIE (2005): "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.

ZOU, H. AND H. H. ZHANG (2009): "On the Adaptive Elastic-net with A Diverging Number of Parameters," *Annals of Statistics*, 37, 1733–1751.

APPENDIX

A Related Proofs

In this section, we provide complete proofs on the two theorems and Corollary 3 presented in the main text. We start with some necessary lemmas.

A.1 Necessary Lemmas

We first show some lemmas that are essential for proving the theorems and propositions in the main text.

Lemma 1 *(Rao, 1973) For a $p \times p$ positive-definite and symmetric matrix U and a p-dimensional vector u,*

$$\sup_z \frac{(z^\top u)^2}{z^\top U z} = u^\top U^{-1} u. \tag{A1}$$

Lemma 2 *(Zhang, 2021) Assume that $\mathbb{E}(e_t) = 0$ and $\mathbb{E}(e_t^4)$ exists. Let $\sigma^2 = \mathbb{E}(e_t^2)$ and $\kappa = E(e_t^4) - 3\sigma^4$. For any two $T \times T$ square matrices O_1 and O_2,*

$$\begin{aligned} \mathbb{E}(e^\top O_1 e e^\top O_2 e) &= \sigma^4 \left\{ trace(O_1) trace(O_2) + trace\left(O_1 O_2 + O_1^\top O_2\right) \right\} \\ &\quad + \kappa \times trace(O_1 * O_2), \end{aligned} \tag{A2}$$

where $*$ denotes the Hadamard product, that is, the ij^{th} component of $O_1 * O_2$ equals the product of the ij^{th} components of O_1 and O_2.

Lemma 3 *(Zhang, 2010; Gao et al., 2019) Let*

$$\widetilde{w} = argmin_{w \in \mathcal{W}} \left\{ L_T(w) + a_T(w) + b_T \right\},$$

where $a_T(w)$ is a term related to w and b_T is a term unrelated to w. If

$$\sup_{w \in \mathcal{W}} |a_T(w)| / R_T(w) = o_p(1), \ \sup_{w \in \mathcal{W}} |R_T(w) - L_T(w)| / R_T(w) = o_p(1),$$

and there exists a constant c and a positive integer T^* so that when $T \geq T^*$, $inf_{w \in \mathcal{W}} R_T(w) \geq c > 0$ almost surely, then $L_T(\widetilde{w}) / inf_{w \in \mathcal{W}} L_T(w) \to 1$ in probability.

Lemma 4 *(Zhang, 2021) For any $T_1 \times T_2$ matrices B_1 and B_2,*

$$\zeta_{\max}(B_1 B_2) \leq \zeta_{\max}(B_1) \zeta_{\max}(B_2), \tag{A3}$$

and

$$\zeta_{\max}(B_1 + B_2) \leq \zeta_{\max}(B_1) + \zeta_{\max}(B_2). \tag{A4}$$

A.2 Proofs of Theorems

Building upon the foundational lemmas presented in Appendix A.1, we proceed to provide comprehensive proofs for Theorems 1 and 2. While the proofs presented below draw inspiration from the work of Zhang (2021), which primarily delves into least squares estimation, our own contribution surpasses these confines. Our method's enhanced generality empowers its application across diverse contexts and complexities.

Proof of Theorem 1. Let $A(w) = I_T - P(w)$. Under Condition C.2, we have that

$$\begin{aligned} R_T(w) &= \mathbb{E}\{L_T(w)|X\} = \mathbb{E}\{\|\widehat{\mu}(w) - \mu\|^2|X\} = \mathbb{E}\{\|P(w)y - \mu\|^2|X\} \\ &= \mathbb{E}\{\|P(w)\mu - \mu + P(w)e\|^2|X\} = \|A(w)\mu\|^2 + \sigma^2 \text{trace}\left\{P(w)P(w)^{\mathrm{T}}\right\}. \end{aligned} \tag{A5}$$

It can be seen that

$$\begin{aligned} C_1(w) - L_T(w) &= \|\widehat{\mu}(w) - y\|^2 + 2\sigma^2 \text{trace}\{P(w)\} - \|\widehat{\mu}(w) - \mu\|^2 \\ &= 2e^\top A(w)\mu + 2\sigma^2 \text{trace}\{P(w)\} - 2e^\top P(w)e + \|e\|^2, \end{aligned}$$

where the last term is unrelated to w, and

$$\begin{aligned} R_T(w) - L_T(w) &= \|A(w)\mu\|^2 + \sigma^2 \text{trace}\left\{P(w)P(w)^{\mathrm{T}}\right\} - \|\widehat{\mu}(w) - \mu\|^2 \\ &= 2e^\top P(w)^\top A(w)\mu + \sigma^2 \text{trace}\{P(w)P(w)^\top\} - e^\top P(w)^\top P(w)e. \end{aligned}$$

In addition, Condition C.7 implies that there exists a constant c and a positive integer T^* so that when $T \geq T^*$, $\xi_T \geq c > 0$ almost surely. Hence from Lemma 3, if we intend to prove Theorem 1, it is sufficient to verify that

$$\sup_{w \in \mathcal{W}} |e^\top A(w)\mu| / R_T(w) = o_p(1), \tag{A6}$$

$$\sup_{w \in \mathcal{W}} |e^\top P(w)^\top A(w)\mu| / R_T(w) = o_p(1), \tag{A7}$$

$$\sup_{w \in \mathcal{W}} |\sigma^2 \text{trace}\{P(w)\} - e^\top P(w)e| / R_T(w) = o_p(1), \tag{A8}$$

$$\sup_{w \in \mathcal{W}} |\sigma^2 \text{trace}\{P(w)^\top P(w)\} - e^\top P(w)^\top P(w)e| / R_T(w) = o_p(1). \tag{A9}$$

For convenience, in all the proofs we assume X_t to be non-stochastic instead of stochastic. This alternative assumption will not invalidate our proof, because all of our technical assumptions concerning X_t hold almost surely.

Proof of (A6): Let $A_{(m)} = I_T - P_{(m)}$ and $\Phi = (\mu^\top A_{(m)}^\top A_{(s)}\mu)_{M_T \times M_T}$, which indicates the ms^{th} component of Φ is $\mu^\top A_{(m)}^\top A_{(s)}\mu$, $G_{T \times M_T} = (A_{(1)}\mu, \ldots, A_{(M_T)}\mu)$, $\Psi = \left\{\sigma^2 \text{trace}(P_{(m)}P_{(s)}^\top)\right\}_{M_T \times M_T}$, and

$$\Psi_0 = \sigma^2 \text{diag}(\text{trace}(P_{(1)}P_{(1)}^\top), \ldots, \text{trace}(P_{(M_T)}P_{(M_T)}^\top)).$$

Therefore $\Phi = G^\top G$. For any $w \in \mathcal{W}$,

$$w^\top \Psi_0 w \leq w^\top \Psi w, \tag{A10}$$

because $w_m \geq 0$, $w_s \geq 0$ for any $m, s \in \{1, \ldots, M_T\}$ and $\text{trace}(P_{(m)}P_{(s)}^\top) \geq 0$ by Condition C.4. In

addition, (A5) implies

$$\begin{aligned} R_T(w) &= \mathbb{E}\|P(w)\mu - \mu + P(w)e\|^2 \\ &= \|A(w)\mu\|^2 + \sigma^2 \text{trace}\left\{P(w)P(w)^{\mathrm{T}}\right\} \\ &= w^\top(\Phi + \Psi)w \\ &\geq w^\top(\Phi + \Psi_0)w, \end{aligned} \tag{A11}$$

where the last step is from (A10). We also have

$$\Phi + \Psi_0 > 0, \tag{A12}$$

because $\Phi = G^\top G$ and $\Psi_0 > 0$ by Condition C.4. Let $\rho = (e^\top A_{(1)}\mu, \ldots, e^\top A_{(M_T)}\mu)^\top$. It is straightforward to demonstrate that

$$\mathbb{E}(\rho) = 0 \tag{A13}$$

and

$$\text{var}(\rho) = \mathbb{E}(\rho\rho^\top) = \mathbb{E}\left\{(e^\top A_{(m)}\mu\mu^\top A_{(s)}e)_{M_T \times M_T}\right\} = \sigma^2\Phi. \tag{A14}$$

It can be seen that

$$\begin{aligned} \sup_{w\in\mathcal{W}} \frac{(e^\top A(w)\mu)^2}{R_T^2(w)} &= \sup_{w\in\mathcal{W}} \frac{(\sum_{m=1}^{M_T} w_m e^\top A_{(m)}\mu)^2}{R_T^2(w)} \\ &= \sup_{w\in\mathcal{W}} \frac{(w^\top\rho)^2}{R_T^2(w)} \\ &\leq \sup_{w\in\mathcal{W}} \frac{(w^\top\rho)^2}{w^\top(\Phi+\Psi_0)w} \sup_{w\in\mathcal{W}} \frac{1}{R_T(w)} \\ &\leq \xi_T^{-1}\rho^\top(\Phi+\Psi_0)^{-1}\rho, \end{aligned} \tag{A15}$$

where the third step is from (A11) and the last step is from (A12) and Lemma 1. By Markov Inequality, we can infer that for any $\delta > 0$,

$$\begin{aligned} & \Pr\left\{\xi_T^{-1}\rho^\top(\Phi+\Psi_0)^{-1}\rho > \delta\right\} \\ \leq\ & \delta^{-1}\xi_T^{-1}\mathbb{E}\left\{\rho^\top(\Phi+\Psi_0)^{-1}\rho\right\} \\ =\ & \delta^{-1}\xi_T^{-1}\sigma^2\text{trace}\left\{(\Phi+\Psi_0)^{-1}\Phi\right\} \\ \leq\ & \delta^{-1}\xi_T^{-1}\sigma^2\text{trace}\left\{(\Phi+\Psi_0)^{-1}\Phi + \Psi_0^{1/2}(\Phi+\Psi_0)^{-1}\Psi_0^{1/2}\right\} \\ =\ & \delta^{-1}\xi_T^{-1}\sigma^2 M_T, \end{aligned} \tag{A16}$$

where the second step is from (A13) and (A14). Combining (A15), (A16) and Condition C.7, we obtain (A6).

Proof of (A7): First, we have that

$$\mathbb{E}(G^\top P_{(m)}^\top e) = 0, \tag{A17}$$

and

$$\text{var}(G^\top P_{(m)}^\top e) \quad = \quad \mathbb{E}(G^\top P_{(m)}^\top e e^\top P_{(m)} G) = \sigma^2 \text{trace}(P_{(m)} G G^\top P_{(m)}^\top). \tag{A18}$$

It can be seen that

$$\begin{aligned}
& \left\{ \sup_{w \in \mathcal{W}} \frac{|e^\top P(w)^\top A(w) \mu|}{R_T(w)} \right\}^2 \\
= & \left\{ \sup_{w \in \mathcal{W}} \frac{|\sum_{m=1}^{M_T} w_m e^\top P_{(m)}^\top A(w) \mu|}{R_T(w)} \right\}^2 \\
\leq & \left\{ \sup_{w \in \mathcal{W}} \frac{\sum_{m=1}^{M_T} w_m |e^\top P_{(m)} A(w) \mu|}{R_T(w)} \right\}^2 \\
\leq & \left\{ \sup_m \sup_{w \in \mathcal{W}} \frac{|e^\top P_{(m)}^\top A(w) \mu|}{R_T(w)} \right\}^2 \\
= & \sup_m \sup_{w \in \mathcal{W}} \frac{(e^\top P_{(m)}^\top G w)^2}{R_T^2(w)} \\
\leq & \sup_m \sup_{w \in \mathcal{W}} \frac{(e^\top P_{(m)}^\top G w)^2}{w^\top (\Phi + \Psi_0) w} \sup_{w \in \mathcal{W}} \frac{1}{R_T(w)} \\
\leq & \xi_T^{-1} \sup_m e^\top P_{(m)}^\top G (\Phi + \Psi_0)^{-1} G^\top P_{(m)} e,
\end{aligned} \tag{A19}$$

where the third step is from $\sum_{m=1}^{M_T} w_m = 1$ and $w_m \geq 0$, the fifth step is from (A11), and the last step is from (A12) and Lemma 1. By Markov Inequality, we can prove that for any $\delta > 0$,

$$\begin{aligned}
& \Pr \left\{ \xi_T^{-1} \sup_m e^\top P_{(m)}^\top G (\Phi + \Psi_0)^{-1} G^\top P_{(m)} e > \delta \right\} \\
\leq & \sum_{m=1}^{M_T} \Pr \left\{ \xi_T^{-1} e^\top P_{(m)}^\top G (\Phi + \Psi_0)^{-1} G^\top P_{(m)} e > \delta \right\} \\
\leq & \sum_{m=1}^{M_T} \delta^{-1} \xi_T^{-1} \mathbb{E} \left\{ e^\top P_{(m)}^\top G (\Phi + \Psi_0)^{-1} G^\top P_{(m)} e \right\} \\
= & \sum_{m=1}^{M_T} \delta^{-1} \xi_T^{-1} \sigma^2 \text{trace} \left\{ P_{(m)}^\top G (\Phi + \Psi_0)^{-1} G^\top P_{(m)} \right\} \\
= & \sum_{m=1}^{M_T} \delta^{-1} \xi_T^{-1} \sigma^2 \text{trace} \left\{ (\Phi + \Psi_0)^{-1/2} G^\top P_{(m)} P_{(m)}^\top G (\Phi + \Psi_0)^{-1/2} \right\} \\
\leq & \sum_{m=1}^{M_T} \delta^{-1} \xi_T^{-1} \sigma^2 \zeta_{\max}(P_{(m)} P_{(m)}^\top) \text{trace} \left\{ (\Phi + \Psi_0)^{-1/2} G^\top G (\Phi + \Psi_0)^{-1/2} \right\} \\
\leq & \max_m \zeta_{\max}(P_{(m)} P_{(m)}^\top) \sum_{m=1}^{M_T} \delta^{-1} \xi_T^{-1} \sigma^2 \text{trace} \left\{ (\Phi + \Psi_0)^{-1/2} \Phi (\Phi + \Psi_0)^{-1/2} \right\} \\
\leq & \max_m \zeta_{\max}(P_{(m)} P_{(m)}^\top) \sum_{m=1}^{M_T} \delta^{-1} \xi_T^{-1} M_T \sigma^2 \\
= & \max_m \zeta_{\max}(P_{(m)} P_{(m)}^\top) \delta^{-1} \xi_T^{-1} M_T^2 \sigma^2,
\end{aligned} \tag{A20}$$

where the third step is from (A17)-(A18) and the sixth step is from that $G^\top G = \Phi$. Combining (A19), (A20) and Conditions C.7 and C.5, we obtain (A7).

Proof of (A8): Let $\tau = \{\sigma^2 \text{trace}(P_{(1)}) - e^\top P_{(1)} e, \ldots, \sigma^2 \text{trace}(P_{(M_T)}) - e^\top P_{(M_T)} e\}^\top$. Then we have that

$$\mathbb{E}(\tau) = 0, \tag{A21}$$

and

$$\begin{aligned}
\text{var}(\tau) &= \mathbb{E}(\tau\tau^\top) \\
&= \mathbb{E}\left[\left\{(\sigma^2 \text{trace}(P_{(m)}) - e^\top P_{(m)} e)(\sigma^2 \text{trace}(P_{(s)}) - e^\top P_{(s)} e)\right\}_{M_T \times M_T}\right] \\
&= \mathbb{E}\left[\left\{e^\top P_{(m)} e e^\top P_{(s)} e - \sigma^4 \text{trace}(P_{(m)}) \text{trace}(P_{(s)})\right\}_{M_T \times M_T}\right] \\
&= \left\{\sigma^4 \text{trace}(P_{(m)} P_{(s)} + P_{(m)}^\top P_{(s)}) + \kappa \text{trace}(P_{(m)} * P_{(s)})\right\}_{M_T \times M_T},
\end{aligned} \tag{A22}$$

where the last step is from Lemma 2. To derive (A22), we also rely on Condition C.2, particularly for the steps regarding κ.[42] Moreover, we can show that

$$\begin{aligned}
\sup_{w \in \mathcal{W}} \frac{\{\sigma^2 \text{trace}\{P(w)\} - e^\top P(w) e\}^2}{R_T^2(w)} &= \sup_{w \in \mathcal{W}} \frac{(w^\top \tau)^2}{R_T^2(w)} \\
&\leq \xi_T^{-1} \sup_{w \in \mathcal{W}} \frac{(w^\top \tau)^2}{w^\top \Psi_0 w} \\
&\leq \xi_T^{-1} \tau^\top \Psi_0^{-1} \tau,
\end{aligned} \tag{A23}$$

where the second step is from (A11) and the last step is from Lemma 1. By Markov Inequality, we show that for any $\delta > 0$,

$$\begin{aligned}
& \Pr\left\{\xi_T^{-1} \tau^\top \Psi_0^{-1} \tau > \delta\right\} \\
\leq\ & \delta^{-1} \xi_T^{-1} \mathbb{E}(\tau^\top \Psi_0^{-1} \tau) \\
=\ & \delta^{-1} \xi_T^{-1} \text{trace}\left[\Psi_0^{-1}\left\{\sigma^4 \text{trace}(P_{(m)} P_{(s)} + P_{(m)}^\top P_{(s)}) + \kappa \text{trace}(P_{(m)} * P_{(s)})\right\}_{M_T \times M_T}\right] \\
=\ & \delta^{-1} \xi_T^{-1} \text{trace}\left[\Psi_0^{-1} \text{diag}\left\{\sigma^4 \text{trace}(P_{(m)}^2 + P_{(m)}^\top P_{(m)}) + \kappa \text{trace}(P_{(m)} * P_{(m)}); m = 1, \ldots, M_T\right\}\right] \\
\leq\ & \delta^{-1} \xi_T^{-1} \text{trace}\left[\Psi_0^{-1} \text{diag}\left\{\sigma^4 \text{trace}(P_{(m)}^\top P_{(m)}) + (\sigma^4 + \kappa) \text{trace}(P_{(m)}^2); m = 1, \ldots, M_T\right\}\right] \\
=\ & \sigma^2 \delta^{-1} \xi_T^{-1} \text{trace}\left[\Psi_0^{-1} \Psi_0\right] + (\sigma^4 + \kappa) \delta^{-1} \xi_T^{-1} \text{trace}\left[\Psi_0^{-1} \text{diag}\left\{\text{trace}(P_{(m)}^2); m = 1, \ldots, M_T\right\}\right] \\
=\ & O(\xi_T^{-1} M_T),
\end{aligned} \tag{A24}$$

where diag $\{b_m; m = 1, \ldots, M_T\}$ denotes a diagonal matrix with the m^{th} component being b_m, the second step is from (A21) and (A22), the fourth step is derived from the fact that Ψ_0^{-1} is a diagonal matrix, and the fifth step is established owing to the finding that for any square matrix O,

$$\text{trace}(O * O) \leq \text{trace}(O^\top O). \tag{A25}$$

[42] For the sake of brevity, we do not repeatedly list Condition C.2 here or elsewhere in the Appendix when it is utilized.

Combining (A23), (A24) and Condition C.7, we obtain (A8).

Proof of (A9): For $j \in \{1, \ldots, M_T\}$, let

$$\nu_{(j)} = (\sigma^2 \text{trace}(P_{(j)}^\top P_{(1)}) - e^\top P_{(j)}^\top P_{(1)} e, \ldots, \sigma^2 \text{trace}(P_{(j)}^\top P_{(M_T)}) - e^\top P_{(j)}^\top P_{(M_T)} e)^\top.$$

Thus we have that

$$\mathbb{E}(\nu_{(j)}) = 0, \tag{A26}$$

and

$$\begin{aligned}
& \text{var}(\nu_{(j)}) \\
= \; & \mathbb{E}(\nu_{(j)} \nu_{(j)}^\top) \\
= \; & \mathbb{E}\left[\left\{ (\sigma^2 \text{trace}(P_{(j)}^\top P_{(m)}) - e^\top P_{(j)}^\top P_{(m)} e)(\sigma^2 \text{trace}(P_{(j)}^\top P_{(s)}) - e^\top P_{(j)}^\top P_{(s)} e) \right\}_{M_T \times M_T} \right] \\
= \; & \mathbb{E}\left[\left\{ e^\top P_{(j)}^\top P_{(m)} e e^\top P_{(j)}^\top P_{(s)} e - \sigma^4 \text{trace}(P_{(j)}^\top P_{(m)}) \text{trace}(P_{(j)}^\top P_{(s)}) \right\}_{M_T \times M_T} \right] \\
= \; & \Big[\sigma^4 \text{trace}(P_{(j)}^\top P_{(m)} P_{(j)}^\top P_{(s)}) + \sigma^4 \text{trace}(P_{(m)}^\top P_{(j)} P_{(j)}^\top P_{(s)}) \\
& + \kappa \text{trace}\left\{ (P_{(j)}^\top P_{(m)}) * (P_{(j)}^\top P_{(s)}) \right\} \Big]_{M_T \times M_T},
\end{aligned} \tag{A27}$$

where the last step is from Lemma 2. By Condition C.5, we can infer

$$\text{trace}(P_{(m)}^\top P_{(j)} P_{(j)}^\top P_{(m)}) \leq \zeta_{\max}(P_{(j)} P_{(j)}^\top) \text{trace}(P_{(m)}^\top P_{(m)}) \leq c_1 \text{trace}(P_{(m)}^\top P_{(m)}). \tag{A28}$$

It is seen that

$$\begin{aligned}
& \left\{ \sup_{w \in \mathcal{W}} \frac{|\sigma^2 \text{trace}\{P(w)^\top P(w)\} - e^\top P(w)^\top P(w) e|}{R_T(w)} \right\}^2 \\
= \; & \left\{ \sup_{w \in \mathcal{W}} \frac{|\sum_{j=1}^{M_T} w_j \sigma^2 \text{trace}\{P_{(j)}^\top P(w)\} - \sum_{j=1}^{M_T} w_j e^\top P_{(j)}^\top P(w) e|}{R_T(w)} \right\}^2 \\
\leq \; & \left\{ \sup_{w \in \mathcal{W}} \frac{\sum_{j=1}^{M_T} w_j |\sigma^2 \text{trace}\{P_{(j)}^\top P(w)\} - e^\top P_{(j)}^\top P(w) e|}{R_T(w)} \right\}^2 \\
\leq \; & \left\{ \sup_j \sup_{w \in \mathcal{W}} \frac{|\sigma^2 \text{trace}\{P_{(j)}^\top P(w)\} - e^\top P_{(j)}^\top P(w) e|}{R_T(w)} \right\}^2 \\
= \; & \sup_j \sup_{w \in \mathcal{W}} \frac{\left\{ \sigma^2 \text{trace}\{P_{(j)}^\top P(w)\} - e^\top P_{(j)}^\top P(w) e \right\}^2}{R_T^2(w)} \\
= \; & \sup_j \sup_{w \in \mathcal{W}} \frac{\left(w^\top \nu_{(j)} \right)^2}{R_T^2(w)}
\end{aligned}$$

$$
\begin{aligned}
&\leq \xi_T^{-1} \sup_j \sup_{w \in \mathcal{W}} \frac{(w^\top \nu_{(j)})^2}{w^\top \Psi_0 w} \\
&\leq \xi_T^{-1} \sup_j \nu_{(j)}^\top \Psi_0^{-1} \nu_{(j)} \\
&\leq \xi_T^{-1} \sum_{j=1}^{M_T} \nu_{(j)}^\top \Psi_0^{-1} \nu_{(j)}, \qquad \text{(A29)}
\end{aligned}
$$

where the sixth step is from (A11) and the seventh step is from Lemma 1. By Markov Inequality, it is clear that for any $\delta > 0$,

$$
\begin{aligned}
&\Pr\left\{\xi_T^{-1} \textstyle\sum_{j=1}^{M_T} \nu_{(j)}^\top \Psi_0^{-1} \nu_{(j)} > \delta\right\} \\
\leq\ & \delta^{-1} \xi_T^{-1} \textstyle\sum_{j=1}^{M_T} \mathbb{E}\left\{\nu_{(j)}^\top \Psi_0^{-1} \nu_{(j)}\right\} \\
=\ & \delta^{-1} \xi_T^{-1} \textstyle\sum_{j=1}^{M_T} \text{trace}\Big[\Psi_0^{-1}\Big\{\sigma^4 \text{trace}(P_{(j)}^\top P_{(m)} P_{(j)}^\top P_{(s)}) + \sigma^4 \text{trace}(P_{(m)}^\top P_{(j)} P_{(j)}^\top P_{(s)}) \\
&\qquad + \kappa \text{trace}((P_{(j)}^\top P_{(m)}) * (P_{(j)}^\top P_{(s)}))\Big\}_{M_T \times M_T}\Big] \\
=\ & \delta^{-1} \xi_T^{-1} \sigma^4 \textstyle\sum_{j=1}^{M_T} \text{trace}\Big[\Psi_0^{-1} \text{diag}\Big\{\text{trace}(P_{(j)}^\top P_{(m)} P_{(j)}^\top P_{(m)}) \\
&\qquad + \text{trace}(P_{(m)}^\top P_{(j)} P_{(j)}^\top P_{(m)}); m = 1, \ldots, M_T\Big\}\Big] \\
&+ \delta^{-1} \xi_T^{-1} \kappa \textstyle\sum_{j=1}^{M_T} \text{trace}\Big[\Psi_0^{-1} \text{diag}\Big\{\text{trace}((P_{(j)}^\top P_{(m)}) * (P_{(j)}^\top P_{(m)})); m = 1, \ldots, M_T\Big\}\Big] \\
=\ & \delta^{-1} \xi_T^{-1} \sigma^4 \textstyle\sum_{j=1}^{M_T} \text{trace}\Big[\Psi_0^{-1} \text{diag}\Big\{\text{trace}(P_{(j)}^\top P_{(m)} P_{(j)}^\top P_{(m)}) \\
&\qquad + \text{trace}(P_{(m)}^\top P_{(j)} P_{(j)}^\top P_{(m)}); m = 1, \ldots, M_T\Big\}\Big] \\
&+ \delta^{-1} \xi_T^{-1} \kappa \textstyle\sum_{j=1}^{M_T} \text{trace}\Big[\Psi_0^{-1} \text{diag}\Big\{\text{trace}(P_{(m)}^\top P_{(j)} P_{(j)}^\top P_{(m)}); m = 1, \ldots, M_T\Big\}\Big] \\
\leq\ & \delta^{-1} \xi_T^{-1} (c_2 + c_1) \sigma^4 \textstyle\sum_{j=1}^{M_T} \text{trace}\left[\Psi_0^{-1} \Psi_0 \sigma^{-2}\right] + \delta^{-1} \xi_T^{-1} \kappa c_1 \sum_{j=1}^{M_T} \text{trace}\left[\Psi_0^{-1} \Psi_0 \sigma^{-2}\right] \\
=\ & \delta^{-1} \xi_T^{-1} M_T^2 (\sigma^4 (c_1 + c_2) + c_1 \kappa) \sigma^{-2}, \qquad \text{(A30)}
\end{aligned}
$$

where the second step is from (A26)-(A27), the fourth step is from (A25), and the fifth step is from (A28) and Condition C.6. Based on (A29), (A30) and Condition C.7, we can prove (A9). So far we have established (A6)-(A9), which are sufficient for validating Equation (24) in Theorem 1.

It is seen that

$$
C_1'(w) = C_1(w) + 2\text{trace}\{P(w)\}(\widehat{\sigma}^2(w) - \sigma^2).
$$

Therefore from Lemma 3, in order to prove (25), we only need to verify that

$$
\sup_{w \in \mathcal{W}} \frac{|\text{trace}\{P(w)\}\{\widehat{\sigma}^2(w) - \sigma^2\}|}{R_T(w)} = o_p(1). \qquad \text{(A31)}
$$

Drawing from Section 2.3.4 of Zhang (2010), we imply that

$$
\sup_{w \in \mathcal{W}} \frac{|\text{trace}\{P(w)\}\{\widehat{\sigma}^2(w) - \sigma^2\}|}{R_T(w)}
$$

$$
\begin{aligned}
&= \sup_{w\in\mathcal{W}} \frac{|\text{trace}\{P(w)\}\{\|A(w)y\|^2/T-\sigma^2\}|}{R_T(w)} \\
&= \sup_{w\in\mathcal{W}} \frac{|\text{trace}\{P(w)\}\{\|\hat{\mu}(w)-\mu\|^2+2\mu^\top A^\top(w)e-2e^\top P(w)e+\|e\|^2-T\sigma^2\}|}{TR_T(w)} \\
&\le \sup_{w\in\mathcal{W}} \frac{L_T(w)}{R_T(w)} \sup_{w\in\mathcal{W}} \frac{|\text{trace}\{P(w)\}|}{T} + \sup_{w\in\mathcal{W}} \frac{2\left|\mu^\top A^\top(w)e\right|}{R_T(w)} \sup_{w\in\mathcal{W}} \frac{|\text{trace}\{P(w)\}|}{T} \\
&\quad + \frac{\left|\|e\|^2-\sigma^2T\right|}{T^{1/2}} \sup_{w\in\mathcal{W}} \frac{1}{R_T(w)} \sup_{w\in\mathcal{W}} \frac{|\text{trace}\{P(w)\}|}{T^{1/2}} \\
&\quad + \sup_{w\in\mathcal{W}} \frac{2\left|e^\top P(w)e-\sigma^2\text{trace}\{P(w)\}\right|}{R_T(w)} \sup_{w\in\mathcal{W}} \frac{|\text{trace}\{P(w)\}|}{T} \\
&\quad + 2\sigma^2 \sup_{w\in\mathcal{W}} \frac{1}{R_T(w)} \sup_{w\in\mathcal{W}} \frac{\text{trace}^2\{P(w)\}}{T}. \qquad \text{(A32)}
\end{aligned}
$$

By Condition C.2, we have that

$$
\frac{\left|\|e\|^2-\sigma^2T\right|}{T^{1/2}} = O_p(1). \qquad \text{(A33)}
$$

Hence, (A31) can be obtained from the previous proofs and Conditions C.3 and C.7. The above discussion concludes the proof of Theorem 1. ∎

Proof of Theorem 2. Since the error term e in this case is assumed to be heteroskedastic with a covariance matrix Ω, the transformed $e^* = \Omega^{-1/2}e$ is known to be homoskedastic. Incorporating the above transformation with the proof of Theorem 1, we have that

$$
\sup_{w\in\mathcal{W}} \frac{\left|C_2(w)-L_T(w)-\|e\|^2\right|}{R_T(w)} = o_p(1) \text{ and } \sup_{w\in\mathcal{W}} \frac{|R_T(w)-L_T(w)|}{R_T(w)} = o_p(1). \qquad \text{(A34)}
$$

Equation (26) is thus proved.

It is seen that

$$
C_2'(w) = C_2(w) + 2\text{trace}\{P(w)\widehat{\Omega}(w)\} - 2\text{trace}\{P(w)\Omega\}.
$$

Hence from Lemma 3, in order to prove (27), we are only left to verify that

$$
\sup_{w\in\mathcal{W}} [|\text{trace}\{P(w)\widehat{\Omega}(w)\} - \text{trace}\{P(w)\Omega\}|/R_T(w)] = o_p(1). \qquad \text{(A35)}
$$

Let $Q_{(m)} = \text{diag}(\iota_{11}^{(m)},\ldots,\iota_{TT}^{(m)})$ and $Q(w) = \sum_{m=1}^{M_T} w_m Q_{(m)}$. Then we have that

$$
\begin{aligned}
&\sup_{w\in\mathcal{W}} [|\text{trace}\{P(w)\widehat{\Omega}(w)\} - \text{trace}\{P(w)\Omega\}|/R_T(w)] \\
= &\sup_{w\in\mathcal{W}} [|\{y-P(w)y\}^\top Q(w)\{y-P(w)y\} - \text{trace}\{Q(w)\Omega\}|/R_T(w)] \\
= &\sup_{w\in\mathcal{W}} [|\{e+\mu-P(w)y\}^\top Q(w)\{e+\mu-P(w)y\} - \text{trace}\{Q(w)\Omega\}|/R_T(w)]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sup_{w\in\mathcal{W}}[|e^\top Q(w)e - \mathrm{trace}\{Q(w)\Omega\}|/R_T(w)] + 2\sup_{w\in\mathcal{W}}[|e^\top Q(w)\{P(w)y-\mu\}|/R_T(w)] \\
&\quad + \sup_{w\in\mathcal{W}}[|\{P(w)y-\mu\}^\top Q(w)\{P(w)y-\mu\}|/R_T(w)] \\
&\leq \sup_{w\in\mathcal{W}}[|e^\top Q(w)e - \mathrm{trace}\{Q(w)\Omega\}|/R_T(w)] + 2\sup_{w\in\mathcal{W}}[|e^\top Q(w)\{P(w)\mu-\mu\}|/R_T(w)] \\
&\quad + 2\sup_{w\in\mathcal{W}}[|e^\top Q(w)P(w)e - \mathrm{trace}\{Q(w)P(w)\Omega\}|/R_T(w)] \\
&\quad + 2\sup_{w\in\mathcal{W}}[|\mathrm{trace}\{Q(w)P(w)\Omega\}|/R_T(w)] \\
&\quad + \sup_{w\in\mathcal{W}}[|\{P(w)y-\mu\}^\top Q(w)\{P(w)y-\mu\}|/R_T(w)] \\
&\equiv \Xi_1 + 2\Xi_2 + 2\Xi_3 + 2\Xi_4 + \Xi_5. \qquad \text{(A36)}
\end{aligned}
$$

Define $\iota = \max_m \max_t \iota_{tt}^{(m)}$. Using Conditions C.5 and C.8, Chebyshev's inequality and Theorem 2 of Whittle (1960), we obtain that, for any $\delta > 0$,

$$
\begin{aligned}
\Pr(\Xi_1 > \delta) &\leq \textstyle\sum_{m=1}^{M_T} \Pr[|e^\top Q_{(m)}e - \mathrm{trace}(Q_{(m)}\Omega)| > \delta\xi_T] \\
&\leq \delta^{-2}\xi_T^{-2}\textstyle\sum_{m=1}^{M_T}\mathbb{E}\{e^\top Q_{(m)}e - \mathrm{trace}(Q_{(m)}\Omega)\}^2 \\
&= O\left\{\xi_T^{-2}\textstyle\sum_{m=1}^{M_T}\mathrm{trace}\{\Omega^{1/2}Q_{(m)}\Omega Q_{(m)}\Omega^{1/2}\}\right\} \\
&= O\left\{\xi_T^{-2}{\zeta_{\max}}^2(\Omega)T\iota^2 M_T\right\} \\
&= O(\xi_T^{-2}M_T\iota^2 T), \qquad \text{(A37)}
\end{aligned}
$$

and

$$
\begin{aligned}
\Pr(\Xi_3 > \delta) &\leq \textstyle\sum_{m=1}^{M_T}\Pr\{|e^\top Q_{(m)}P_{(m)}e - \mathrm{trace}(Q_{(m)}P_{(m)}\Omega)| > \delta\xi_T\} \\
&\leq \delta^{-2}\xi_T^{-2}\textstyle\sum_{m=1}^{M_T}\mathbb{E}[e^\top Q_{(m)}P_{(m)}e - \mathrm{trace}(Q_{(m)}P_{(m)}\Omega)]^2 \\
&= O\left\{\xi_T^{-2}\textstyle\sum_{m=1}^{M_T}\mathrm{trace}\{\Omega^{1/2}Q_{(m)}P_{(m)}\Omega P_{(m)}^\top Q_{(m)}\Omega^{1/2}\}\right\} \\
&= O\left\{\xi_T^{-2}{\zeta_{\max}}^2(\Omega)T\iota^2 M_T \max_{1\leq m\leq M_T}\zeta_{\max}(P_{(m)}P_{(m)}^\top)\right\} \\
&= O(\xi_T^{-2}T\iota^2 M_T \max_{1\leq m\leq M_T}\zeta_{\max}(P_{(m)}P_{(m)}^\top)) \\
&= O(\xi_T^{-2}T\iota^2 M_T). \qquad \text{(A38)}
\end{aligned}
$$

It follows from (A37)-(A38) and Condition C.9 that $\Xi_1 + \Xi_3 = o_p(1)$. From Condition C.8 and the second part of (A34), we have

$$
\begin{aligned}
\Xi_2 &\leq \sup_{w\in\mathcal{W}}\{\|e\|^2\iota^2\|P(w)\mu-\mu\|^2/R_T^2(w)\}^{1/2} \\
&\leq \|e\|\iota\xi_T^{-1/2} = O(T^{1/2}\xi_T^{-1/2}\iota),
\end{aligned}
$$

$$
\begin{aligned}
\Xi_4 &\leq \xi_T^{-1}\iota\zeta_{\max}(\Omega)\sup_{w\in\mathcal{W}}[\mathrm{trace}\{P(w)\}] \\
&\leq \xi_T^{-1}\iota\zeta_{\max}(\Omega)\sup_m\{\mathrm{trace}(P_{(m)})\}
\end{aligned}
$$

$$
\begin{aligned}
&= O(\xi_T^{-1}\iota \sup_m \{\mathrm{trace}(P_{(m)})\}) \\
&= O(\xi_T^{-1}\iota^2 T), \qquad \text{(A39)}
\end{aligned}
$$

and

$$
\begin{aligned}
\Xi_5 &\leq \iota \sup_{w\in\mathcal{W}} [\{P(w)y-\mu\}^\top \{P(w)y-\mu\}/R_T(w)] \\
&= \iota \sup_{w\in\mathcal{W}} [L_T(w)/R_T(w)] = O(\iota). \qquad \text{(A40)}
\end{aligned}
$$

Moreover, by Conditions C.7 and C.9, we see that $\Xi_2 + \Xi_4 + \Xi_5 = o_p(1)$. Hence (A35) is proved. This completes the proof of Theorem 2. ∎

A.3 Proof of Corollary 3

In this section, we present the proof of Corollary 3.

Proof of Corollary 3. It is seen that $P_{(m)}^{\mathrm{LSSVR}}$ is symmetric and non-negative definite,

$$
\zeta_{\max}(P_{(m)}^{\mathrm{LSSVR}}) = \zeta_{\max}\{H(H^\top H + \lambda I_T)^{-1}H^\top\} \leq 1, \qquad \text{(A41)}
$$

and

$$
\begin{aligned}
& \mathrm{trace}({P_{(m)}^{\mathrm{LSSVR}}}^\top P_{(m)}^{\mathrm{LSSVR}}) \\
= \ & \mathrm{trace}\{H_{(m)}(H_{(m)}^\top H_{(m)} + \lambda I_T)^{-1}H_{(m)}^\top H_{(m)}(H_{(m)}^\top H_{(m)} + \lambda I_T)^{-1}H_{(m)}^\top\} \\
= \ & \mathrm{trace}\{(H_{(m)}^\top H_{(m)} + \lambda I_T)^{-1}H_{(m)}^\top H_{(m)}(H_{(m)}^\top H_{(m)} + \lambda I_T)^{-1}H_{(m)}^\top H_{(m)}\} \\
= \ & \mathrm{trace}\{(\Psi_{(m)}K_{(m)}\Psi_{(m)}^\top + \lambda I_T)^{-1}\Psi_{(m)}K_{(m)}\Psi_{(m)}^\top(\Psi_{(m)}K_{(m)}\Psi_{(m)}^\top + \lambda I_T)^{-1} \\
& \qquad \times \Psi_{(m)}K_{(m)}\Psi_{(m)}^\top\} \\
= \ & \mathrm{trace}\{(K_{(m)} + \lambda I_T)^{-1}K_{(m)}(K_{(m)} + \lambda I_T)^{-1}K_{(m)}\} \\
= \ & \mathrm{trace}\{(K_{(m)} + \lambda I_T)^{-1}K_{(m)}(K_{(m)} + \lambda I_T)^{-1}K_{(m)}\} \\
= \ & \textstyle\sum_{i=1}^n K_{(m),i}^2(K_{(m),i} + \lambda)^{-2}, \qquad \text{(A42)}
\end{aligned}
$$

where $K_{(m)}$ and $\Psi_{(m)}$ are diagonal matrices containing eigenvalues of $H_{(m)}^\top H_{(m)}$ and the matrices are combined by eigenvectors of $H_{(m)}^\top H_{(m)}$. From Condition (30), we have that there exists a positive constant c such that

$$
\textstyle\sum_{i=1}^n K_{(m),i}^2(K_{(m),i} + \lambda)^{-2} \geq c > 0. \qquad \text{(A43)}
$$

Then, by Theorem 2, Equation (A42) and Equation (A43), Corollary 3 holds. ∎

B Further Details on Prediction Estimators

In this section, we complement the discussion in Section 2 and provide additional details on the considered forecasting estimators, which are also employed as competitive methods in the forecasting exercise. We first review the penalized regression that combines penalty terms with loss functions in a linear framework. We then discuss four commonly employed tree-structured modeling techniques. Since SVR and LSSVR are closely related to our advocated Mallows-type averaging learning, we further show the details about formulation and estimation of SVR and LSSVR in the last subsection.

B.1 Penalized Regression

If we combine a mathematical penalty term with the loss function to be optimized,[43] this brings about the so-called penalized regression which has led to a wide range of applications. Two popular proposals have been made in the literature about how to control the complexity of fitted values through penalty terms:

1. Constrain the sum of absolute values of regression coefficients to be less than some constant C (sometimes called an L_1-penalty); and
2. Constrain the sum of squared regression coefficients to be less than some constant C (sometimes called an L_2-penalty).

In this subsection, we briefly review three widely-applied penalized regressions: ridge regression, the least absolute shrinkage selective operator (LASSO), and the elastic net under the customary setting of $C = 0$.

Ridge regression imposes a constraint on the sum of squared non-intercept coefficients (also known as L_2-penalty), that is,

$$\hat{\beta}_* = \arg\min_{\beta_*} \left[\sum_{t=1}^{T} \left(y_t - \beta_0 - \sum_{i=1}^{p} \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{p} \beta_i^2 \right],$$

where $\beta_* = [\beta_1, ..., \beta_p]^\top$ does not include β_0 and λ is a tuning parameter that determines the severity of penalty. In practice, we either predetermine λ or compute it via certain validation algorithm (for example, five-fold cross-validation).

It follows that the ridge regression estimator is

$$\hat{\beta}_* = \left(X_*^\top X_* + \lambda I \right)^{-1} X_*^\top y, \tag{A44}$$

[43] The penalty imposing greater losses as a mean function becomes more complicated. For greater complexity to be accepted, the fit must be improved by a degree larger than the penalty. Therefore greater complexity has to be worth it. Strategies designed to control the magnitude of coefficients through penalty terms are also called shrinkage or regularization.

where I is a $p \times p$ identity matrix. Note that the matrix X_* is formed by the original regressor matrix X dropping the column of ones for the intercept and β_0 is estimated separately.[44] Once $\hat{\beta}_*$ is obtained, the intercept $\hat{\beta}_0$ is simply the mean of the vector $y - X_*\hat{\beta}_*$. Define the ridge coefficient vector as $\hat{\beta}^{\text{Ridge}} = [\hat{\beta}_0, \hat{\beta}_*^\top]^\top$. The forecast of y_t can be given following Equation (2) from the main text by replacing $\hat{\beta}$ with $\hat{\beta}^{\text{Ridge}}$. That said, the forecast of y_{T+h} is simply $\hat{y}_{T+h} = X_{T+h}^\top \hat{\beta}^{\text{Ridge}}$.

The ridge regression reduces the mean squared error through a trade-off between the prediction bias and the variance. With non-zero λ, the estimator (A44) is clearly biased. However, the reduced variance of the ridge estimates often results in a smaller mean square error than that of the least-squares estimates.

If one adopts the L_1-penalty by restricting the sum of absolute values of the non-intercept coefficients, the corresponding regression procedure is known as LASSO (Tibshirani, 1996). Similar to the ridge regression, LASSO minimizes the following penalized residual sum of squares:

$$\hat{\beta}_* = \arg\min_{\beta_*} \left[\sum_{t=1}^{T} \left(y_t - \beta_0 - \sum_{i=1}^{p} \beta_i x_{it} \right)^2 + \lambda \sum_{i=1}^{p} |\beta_i| \right].$$

Unlike ridge regression, the LASSO penalty leads to a nonlinear estimator for $\hat{\beta}_*$ without an analytical expression. A numerical method via the quadratic programming is needed in this case. Once we estimate $\hat{\beta}_*$, the intercept term $\hat{\beta}_0$ can be computed in the same fashion as in ridge regression. The LASSO coefficients is denoted as $\hat{\beta}^{\text{LASSO}}$. The forecast of y_{T+h} is straightforward: $\hat{y}_{T+h} = X_{T+h}^\top \hat{\beta}^{\text{LASSO}}$.

For the LASSO regression, λ is also a tuning parameter, which could yield the usual least squares estimates with λ equal to zero. As the value of λ increases, the regression coefficients are shrunk towards zero. The LASSO regression is capable of shrinking coefficients to exactly zero without setting $\lambda = \infty$. Therefore, it can be used as a variable selection tool in practice. This concept is illustrated geometrically in James et al. (2017).

Zou and Hastie (2005) pointed out that the LASSO solution paths are unstable when predictors are highly correlated. If there is a group of variables with strong correlations, LASSO is indifferent among various predictor sets. To overcome such limitation, the elastic net is proposed by Zou and Hastie (2005) as an improved version of LASSO. The elastic net is a mixture of ridge regression and LASSO carrying the below penalty

$$\lambda \left[(1-\alpha) \sum_{t=1}^{p} \beta_t^2 + \alpha \sum_{t=1}^{p} |\beta_t| \right], \tag{A45}$$

where $\alpha \in [0,1]$ is called the mixing parameter and λ has the usual interpretation as in ridge and LASSO regressions. The L_1-penalty in (A45) implements variable selection, and the L_2-penalty brings the grouping effect and stabilizes the L_1 solution path. The elastic net includes LASSO and ridge as its special cases when $\alpha = 1$ and $\alpha = 0$, respectively.

[44] By default, the coefficient vector $\hat{\beta}_*$ is computed after standardizing all the predictors to have a mean zero and a standard deviation of one. The model does not include a constant term, and consequently X_* should not contain a column of ones.

B.2 Tree-Structured Learning Methods

This section provides additional descriptions about four major tree-structured learning algorithms used in our exercise. The first method is termed regression tree (RT) proposed by Breiman et al. (1984).[45] Suppose we collect a standard data set $\{y_t, X_t\}_{t=1}^{T}$. Starting from the original data (i.e., the root node), the trick for researchers in applying RT is to find the best split where two stages of searches are usually undertaken. At the first stage, for each predictor all possible binary splits of the predictor values are considered and the best split is determined afterwards.[46] With the best split of each predictor, the best split overall is determined at the second stage, which acts as the winning split to separate the data. Now there are two partitions of the original data. Such partitioning process can be implemented in a recursive fashion until we reach a pre-determined boundary. There are many tuning parameters (also called hyperparameters in the machine learning literature) that need to be decided or calculated beforehand such as the splitting criterion function, the minimum leaf size, the stopping rule, etc.

Data in the terminal nodes (also known as tree leaves) are considered as homogeneous, hence a simple average of the observations of y_l within tree leaf l is used as the fitted value. If there are L tree leaves in total, $\cup_{l=1}^{L} y_l$ is equivalent to the original data y, and $\sum_{l=1}^{L} T_l = T$ with T_l being the number of observations in tree leaf l. To make predictions based on X_{T+h}, we simply drop X_{T+h} down the tree and end up with a specific tree leaf l. The prediction, $\hat{y}_{T+h}$, is measured by $\bar{y}_{t \in l}$, a simple average of the associated observations of the dependent variable at final leaf l.

We next consider bootstrap aggregating decision trees or bagging developed by Breiman (1996). In contrast to the RT method, the bagging (BAG) method involves a training process where the level of training (cycles of learning) is predetermined. The BAG algorithm is summarized as:

1. Take a random sample of size T with replacement (that is, bootstrapping) from the original data. Iteration of the bootstrapping process for B times generates B new training sets. Denote the collection of B bootstrap samples as $\{y_t^{(b)}, X_t^{(b)}\}_{t=1}^{T}$ for $b = 1, ..., B$.
2. Construct a regression tree based on $\{y_t^{(b)}, X_t^{(b)}\}_{t=1}^{T}$.
3. Use the regression tree to make the prediction $\hat{y}_{T+h}^{(b)}$ based on X_{T+h}.
4. Repeat steps (1) to (3) for B times and obtain $\hat{y}_{T+h}^{(b)}$ for each b.
5. Take a simple average of the B forecasts $\hat{y}_{T+h} = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_{T+h}^{(b)}$ as the final forecast.

Besides the conventional tuning parameters for constructing a tree, we also need to determine the value of B for bagging. In most cases, the more bootstrap samples in the training process,

[45] Note that the full name of the method is Classification and Regression Trees (CART), of which classification mostly deals with categorical responses of non-numeric symbols and texts, whereas regression trees concentrate purely on quantitative response variables. Given the numerical nature of our study, we only utilize the regression tree part of CART.

[46] A best split is determined by improvement of a given criterion function, for example, the reduction of SSR. A simple regression of $\{y_t, X_t\}_{t=1}^{T}$ will yield a sum of squared residuals, SSR_0. Suppose we can split the original sample into two sub-samples such that $T = T_1 + T_2$. The RT method finds the best split of a sample in terms of minimizing the sum of squared residuals (SSR) from the sample partitioning. That is, the aggregated SSR values computed from regressions with each sub-sample should obey the relationship: $\text{SSR}_1 + \text{SSR}_2 \leq \text{SSR}_0$.

the better the forecast accuracy. However, a large number of bootstrap samples also imply longer computational time. A balance needs to be reached between accuracy and time constraints.

Random forest (RF) by Breiman (2001) can be regarded as a modification of bagging that reduces the possibility of yielding correlated trees. Similar to bagging, RF also constructs B new trees with bootstrap samples generated from the original data set. But for RF, as each tree is built, we take a random sample (without replacement) of q predictors out of the total p $(q < p)$ predictors before each splitting. The default value for q is set to $\lfloor 1/3q \rfloor$. Such a process is repeated for each node. Briefly speaking, if $q = p$, RF is equivalent to BAG. Eventually, we end up with B trees just as bagging and the final RF forecast is calculated as the simple average of forecasts over all the generated trees.

Both BAG and RF rely on the bootstrap resampling process. Asymptotically, we exploit only 63.2% $(1 - 1/e \approx 63.2\%)$ of the unique observations of $\{y_t, X_t\}_{t=1}^{T}$ with some repeated observations. The remaining observations are called out-of-bag (OOB) observations, which become an ideal test set to evaluate the constructed tree. For each input observation, denoted as X_i, we can find the prediction $\hat{y}_i^{(b)}$ from each bootstrap sample b which does not contain X_i. Theoretically, the number of such predictions should be around $0.37B$. We average these predictions to obtain the OOB averaged prediction $\hat{y}_i^{oob}$. The assessment of bagging and RF performance is based on the OOB error, calculated as $y_i - \hat{y}_i^{oob}$. Finally, we compute its SDFE as our benchmark

$$\text{SDFE} = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(y_i - \hat{y}_i^{oob})^2}$$

to evaluate all the observations and select optimal values for the tuning parameters (e.g., minimum tree leaf size).

The RF method can respond to highly local features of the data, since it depends on a very flexible fitting procedure. This flexibility sometimes comes with a risk of overfitting the data. Therefore boosting is employed as the fourth tree-structured learning method in our paper, which usually serves as a popular alternative ensemble method to random forests. Boosting works by sequentially assigning more weights to the residuals from the prediction of previously grown trees to create the new tree. After a large number of trials, boosting gradually improves its local prediction in areas where it shows poor performance previously. As described in Hastie et al. (2009, Chapter 10), boosting combines the outputs of many weak fitting functions to produce a powerful committee. Boosted trees are typically shallow, with the maximum depth of variable interactions often set to be less than 4 or 5.

In this paper, we consider a simple least squares boosting (LSB) that fits RT ensembles. In line with Hastie et al. (2009, Chapter 10), the LSB method applies a new learning RT at each step to the difference between the observed response and the aggregated prediction of all RT previously grown.[47] More formally, boosting fits an additive basis function expansion that takes the below form

$$f(X) = \sum_{k=1}^{K} \delta_k b(X; \gamma_k),$$

where $b(X; \gamma) \in \mathbb{R}$ are usually simple functions of the multivariate argument X, featured by a set of parameters $\gamma = (\gamma_1, \ldots, \gamma_k)$ for a total of K trees. The associated expansion coefficients are

[47] Many of the boosting methods are designed for classification issues. See, `AdaBoost.M1` by Freund and Schapire (1997) for example.

captured by δ_k for $k = 1, 2, ..., K$. The above models are fit by minimizing a loss function averaged over the training set,

$$\min_{\{\delta_k, \gamma_k\}_1^K} \sum_{t=1}^{T} L\left(y_t, \sum_{k=1}^{K} \delta_k b(x_t; \gamma_k)\right). \tag{A46}$$

For the LSB method, the loss function is represented by squared error loss,

$$L\big(y, f(X)\big) = \big(y - f(X)\big)^2. \tag{A47}$$

B.3 SVR-type Learning Algorithms

Another popular machine learning method that responds to local characteristics of the data is support vector regression (SVR) introduced by Drucker et al. (1996). SVR can be regarded as a regression extension of support vector machine (SVM) to consider real-valued outcome variables.[48] The SVR framework considers the nonlinear formulation (4) from the main text and approximates the nonlinear regression function $f(X_t)$ with a set of basis function $\{h_s(X_t)\}_{s=1}^S$:

$$y_t = f(X_t) + e_t = \sum_{s=1}^{S} \beta_s h_s(X_t) + e_t = \beta^\top h(X_t) + e_t, \tag{A48}$$

where $\beta = [\beta_1, \cdots, \beta_S]^\top$ and $h(\cdot) : \mathbb{R}^p \to \mathbb{R}^S$ is a set of basis functions on X_t.[49] Note that the set of basis functions can be infinite-dimensional. That is, S can go to infinity, and we do not (need to) know the explicit form of the basis functions.

We estimate the coefficients β through the minimization of

$$L(\beta) = \sum_{t=1}^{T} V_e\left(y_t - f(X_t)\right) + \lambda \sum_{s=1}^{S} \beta_s^2, \tag{A49}$$

where the loss function

$$V_\epsilon(r) = \begin{cases} 0 & \text{if } |r| < \epsilon \\ |r| - \epsilon & \text{otherwise} \end{cases}$$

is called an ϵ-insensitive error measure that ignores errors of size less than ϵ. As part of the loss function V_ϵ, the parameter ϵ is usually predetermined. On the other hand, λ is a more traditional regularization parameter that can be estimated by cross-validation.

The minimization problem in (A49) can be modified into

$$\min_{\beta, \xi, \xi^*} J(\beta, \xi, \xi^*) = \lambda \beta^\top \beta + \sum_{t=1}^{T} (\xi_t + \xi_t^*),$$

such that

$$y_t - \beta^\top h(X_t) \le e + \xi_t, \quad \beta^\top h(X_t) - y_t \le e + \xi_t^*, \quad \xi_t, \xi_t^* \ge 0$$

[48] SVM is a supervised learning algorithm that analyzes data for classification applications. The theoretical background is provided in Vapnik (1996). As pointed out by Lehrer and Xie (2022), SVM permits complex nonlinear relationships through transforming the original data into a high dimensional space via a predetermined mapping scheme.

[49] Following Hastie et al. (2009, Chapter 12), we ignore the intercept for simplicity. In practice, this is usually achieved by first standardizing the data for estimation, and then converting the data back for forecasting.

for $t = 1, ..., T$, where ξ_t and ξ_t^* are non-negative slack variables that are chosen to satisfy the above conditions. The Lagrangian for this problem is

$$\begin{aligned}\mathcal{L}(\beta, \xi, \xi^*; \alpha', \alpha^*, \eta', \eta^*) &= \lambda\beta^\top\beta + \sum_{t=1}^{T}(\xi_t + \xi_t^*) - \sum_{t=1}^{T}\alpha_t'(e + \xi_t - y_t + \beta^\top h(X_t)) \\ &\quad - \sum_{t=1}^{T}\alpha_t^*(e + \xi_t^* + y_t - \beta^\top h(X_t)) - \sum_{t=1}^{T}(\eta_t'\xi_t' + \eta_t^*\xi_t^*)\end{aligned}$$

with positive Lagrangian multipliers $\alpha_t', \alpha_t^*, \eta_t', \eta_t^* \geq 0$. Take the first order condition and substitute all the parameters in terms of α' and α^*. The dual problem of the above Lagrangian can be written as

$$\min_{\alpha_t', \alpha_t^*} \epsilon\sum_{t=1}^{T}(\hat{\alpha}_t^* + \hat{\alpha}_t') - \sum_{t=1}^{T} y_t(\hat{\alpha}_t^* - \hat{\alpha}_t') + \sum_{t,t'=1}^{T}(\hat{\alpha}_t^* - \hat{\alpha}_t')(\hat{\alpha}_{t'}^* - \hat{\alpha}_{t'}')h(X_t)^\top h(X_{t'}), \tag{A50}$$

subject to the constraints

$$0 \leq \hat{\alpha}_t^*, \hat{\alpha}_t' \leq \frac{1}{2\lambda}, \quad \sum_{t=1}^{T}(\hat{\alpha}_t^* - \hat{\alpha}_t') = 0, \quad \hat{\alpha}_t'\hat{\alpha}_t^* = 0,$$

for all $t = 1, \ldots, T$. The non-zero values of $\hat{\alpha}_t^* - \hat{\alpha}_t'$ are usually treated as the support vector.

Define $K(X_t, X_{t'}) = h(X_t)^\top h(X_{t'}) \equiv \sum_{s=1}^{S} h_s(X_t)h_s(X_{t'})$ as a kernel function for any input vectors X_t and $X_{t'}$. The solution of Equation (A49) takes the form

$$\hat{f}(x) = \sum_{t=1}^{T}(\hat{\alpha}_t^* - \hat{\alpha}_t')K(x, X_t), \tag{A51}$$

for any input vector x, where $\hat{\alpha}_t^*$ and $\hat{\alpha}_t'$ are the solutions of Equation (A50).

As Equations (A50) and (A51) indicate, no explicit form of the basis function is needed in the calculation process. It is the kernel function that plays a crucial role SVR estimation. In practice, the kernel function is usually predetermined. Common choices of kernel functions are presented below:

$$\begin{aligned}\text{Linear} &: K(x, X_t) = x^\top X_t, \\ \text{Gaussian} &: K(x, X_t) = \exp\left(-\frac{\|x - X_t\|^2}{2\sigma_x^2}\right), \\ \text{Polynomial} &: K(x, X_t) = (\gamma + x^\top X_t)^d,\end{aligned}$$

where σ_x^2, γ, and d are the hyperparameters that can be tuned via cross-validation.

Suykens and Vandewalle (1999) proposed a modification to the classic SVM which leads to solving a set of linear equations under a least-square loss function, henceforth LS SVM. In general, the LS SVM is more computationally efficient than the classic SVM and is capable of dealing with large dataset with high dimensionality. The LS SVM can be implemented to solve for both classification and regression estimation (see Suykens et al. (2002) for an extension to regression applications). In this paper, we mainly focus on LS SVM for regression (LSSVR). Similar to the

minimization problem in (A49), the LSSVR considers minimizing

$$C(\beta) = \sum_{t=1}^{T} (y_t - f(X_t))^2 + \lambda \sum_{s=1}^{S} \beta_s^2. \tag{A52}$$

Compared to the classic SVR formulation, a squared loss function is taken for the error variables in (A52) for LSSVR. In fact, the problem in (A52) can be regarded as a nonparametric ridge regression function formulated in the feature space.

We can construct the Lagrangian equation based on (A52)

$$\mathcal{L}(\beta, \alpha) = C(\beta) - \sum_{t=1}^{T} \alpha_t \left(\beta^\top h(X_t) - y_t \right),$$

where $\alpha = [\alpha_1, ..., \alpha_T]^\top$ are Lagrange multipliers for LSSVR, which can be substituted for β. We solve for α and obtain $\hat{\alpha} = (HH^\top + \gamma I_T)^{-1} y$, where $H = h(X)$ is the implicit basis matrix and $HH^\top$ is the $T \times T$ kernel matrix with the $\{tt'^{th}\}$ element being $K(X_t, X_{t'})$ as in (A51). The resulting LSSVR model for estimation is described by

$$\hat{f}(x) = \sum_{t=1}^{T} \hat{\alpha}_t K(x, X_t),$$

where $\hat{\alpha}_t$ is the estimated Lagrangian multiplier and $K(\cdot, \cdot)$ is the predetermined kernel function that can be linear, Gaussian, or polynomial. Note that when $K(\cdot, \cdot)$ is linear, the LSSVR estimate is identical to the ridge regression discussed in Appendix B.1.

C Derivation of $P^{\mathbf{LSSVR}}(X)$ with Intercept Terms

In line with De Brabanter et al. (2011), if the formulation of LSSVR includes an intercept term β_0 such that

$$y_t = f(X_t) + e_t = \beta_0 + \sum_{s=1}^{S} \beta_s h_s(X_t) + e_t \quad \text{for } t = 1, ..., T, \tag{A53}$$

the optimization problem in LSSVR considers

$$\min_{\beta} \; L(\beta) = \sum_{t=1}^{T} (y_t - f(X_t))^2 + \lambda \sum_{s=1}^{S} \beta_s^2$$

subject to (A53), where $\beta = [\beta_0, ..., \beta_S]^\top = [\beta_0, \beta_*^\top]^\top$. We can construct the following Lagrangian equation

$$\mathcal{L}(\beta, \alpha) = L(\beta) - \sum_{t=1}^{T} \alpha_t \left(\beta_0 + \sum_{s=1}^{S} \beta_s h_s(X_t) - y_t \right),$$

where $\alpha = [\alpha_1, ..., \alpha_T]^\top$ are Lagrange multipliers.

Taking the first-order conditions for optimization and substitute for β_*, we obtain the following solution

$$\begin{bmatrix} 0 & \iota^\top \\ \iota & HH^\top + \lambda I_T \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \tag{A54}$$

where $\iota = [1, ..., 1]^\top$, H is the implicit basis matrix, and $HH^\top$ is the $T \times T$ kernel matrix with the $\{tt'^{th}\}$ element being the kernel function $K(X_t, X_{t'})$. For simplicity, we define

$$\Omega \equiv (HH^\top + \lambda I_T)^{-1}$$

and solve for $\hat{\beta}_0$ and $\hat{\alpha}$ from (A54) such that

$$\begin{aligned} \hat{\alpha} &= \Omega(y - \hat{\beta}_0 \iota) \\ \hat{\beta}_0 &= \iota^\top \Omega y / \iota^\top \Omega \iota. \end{aligned}$$

The resulting LSSVR model now has the form

$$\begin{aligned} \hat{f}(X) &= HH^\top \hat{\alpha} + \hat{\beta}_0 \iota \\ &= HH^\top \Omega(y - \hat{\beta}_0 \iota) + \hat{\beta}_0 \iota \\ &= HH^\top \Omega y + (\iota - HH^\top \Omega \iota)\hat{\beta}_0 \\ &= \left(HH^\top \Omega + \frac{(\iota - HH^\top \Omega \iota)\iota^\top \Omega}{\iota^\top \Omega \iota} \right) y, \\ &= P^{\text{LSSVR}}(X) y, \end{aligned}$$

where

$$P^{\text{LSSVR}}(X) \equiv HH^\top \Omega + \frac{(\iota - HH^\top \Omega \iota)\iota^\top \Omega}{\iota^\top \Omega \iota}. \tag{A55}$$

The no-intercept version of $P^{\text{LSSVR}}(X)$ in (20) can be rewritten as $\Omega HH^\top$. Note that the $T \times T$ matrix $HH^\top \Omega$ is symmetric, since

$$\begin{aligned} HH^\top (HH^\top + \lambda I_T)^{-1} &= HH^\top \left((HH^\top)^{-1} - \lambda (HH^\top)^{-1} (HH^\top + \lambda I_T)^{-1} \right) \\ &= I_T - \lambda (HH^\top + \lambda I_T)^{-1} \end{aligned}$$

following the Woodbury matrix identity. Therefore, the no-intercept version of $P^{\text{LSSVR}}(X)$ is a special case of (A55) that excludes the second term on the right-hand-side of (A55).

D Monte Carlo Simulation

In this section, we conduct Monte Carlo simulation to evaluate the out-of-sample performance of the proposed averaging machine learning methods relative to other competitive estimators. Inspired by Lehrer and Xie (2022), we consider the true DGP of the response variable to follow the below equation:

$$y_t = \sin(x_{1t}) + \cos(x_{2t}) + e_t \quad \text{for } t = 1, ..., T + 1.$$

Suppose that we have access to a set of p predictors $X_t = [x_{1t}, x_{2t}, ..., x_{pt}]^\top$ and therefore $p - 2$ of them are redundant. The exact identification of $p - 2$ redundant variables is unknown to us as what happens usually in reality. All the $\{x_{it}\}_{i=1}^p$ follow $x_{it} \sim i.i.d.\text{N}(0, 4)$ for $i = 1, ..., p$ and the error term e_t can draw from any of the following two distributions to create homoskedasticity and

pure random heteroskedasticity

$$e_t \sim \begin{cases} \mathrm{N}(0,1) & \text{under homoskedasticity,} \\ \mathrm{N}(0, 0.05x_{1t}^2 + 0.01) & \text{under heteroskedasticity.} \end{cases}$$

We generate the data for $t = 1, ..., T+1$ and use T periods of the sample as the training set. Finally, the forecasts of y_{T+1} are made based on the test set of X_{T+1} and the trained forecast methods.

Forecasts of y_{T+1} are calculated from the conventional learning methods and model averaging learning methods: (1) LS; (2) LASSO; (3) RT; (4) BAG; (5) RF; (6) $\mathrm{SVR_L}$; (7) $\mathrm{SVR_G}$; (8) $\mathrm{LSSVR_G}$; (9) $\mathrm{LSSVR_G^{SA}}$; (10) $\mathrm{LSSVR_{G1}^{MA}}$; and (11) $\mathrm{LSSVR_{G2}^{MA}}$.

In this experiment, we set $p = 4$. Alternative values of p have been tried and the results remain the same qualitatively. To facilitate replication, the associated hyperparameters are set to their default values.[50] The major ones include and are not limited to:

1. The penalty coefficient is set to one for LASSO and SVR-type methods;
2. The minimum leaf size is set to one for all the tree-type methods;
3. The learning cycles are assumed to be 100 for all ensemble methods;
4. The number of selected predictors is set at $\lfloor p/3 \rfloor$ for the RF method;
5. $\sigma_x = 1$ for the Gaussian kernel;
6. Candidate model sets are constructed by a full combination of all the included predictors.

For each method, the number of replications is set to $B = 1000$ and a list of forecasts $\hat{y}_{T+1}^{(b)}$ are compared with the actual $y_{T+1}^{(b)}$ for $b = 1, ..., B$. The forecasting performance is assessed by the following two risks:

$$\mathrm{SDFE} = \sqrt{\frac{1}{B}\sum_{b=1}^{B} e_{(b)}^2}, \ \mathrm{MAFE} = \frac{1}{B}\sum_{b=1}^{B} |e_{(b)}|,$$

where $e_{(b)} = y_{T+1}^{(b)} - \hat{y}_{T+1}^{(b)}$ is the forecast error in the b^{th} simulation.

Table A1 reports simulation results for $T = 50$ with the best result under each risk in boldface. The first column reports competitive forecasting strategies, whereas the next two columns and the last two columns correspond to the outcomes under homoskedasticity and heteroskedasticity, respectively.

Several main findings are worth stressing. The least square estimates (henceforth, LS) of a generalized unrestricted model with all the p predictors yield relatively high risks in general. LASSO, $\mathrm{SVR_L}$ and $\mathrm{SVR_G}$ also display poor performance in this case. The performance of RT is disappointing under homoskedasticity. In contrast, its accuracy is improved dramatically under heteroskedasticity. The remaining learning algorithms (i.e., BAG, RF, $\mathrm{LSSVR_G}$ and $\mathrm{LSSVR_G^{SA}}$) perform much better than conventional regressions. Most importantly, we find out that the proposed averaging methods ($\mathrm{LSSVR_{G1}^{MA}}$ and $\mathrm{LSSVR_{G2}^{MA}}$) always improve on its base method $\mathrm{LSSVR_G}$ in terms of yielding lower SDFE and MAFE. Under both homoskedasticity and heteroskedasticity,

[50] We also consider choosing the hyperparameters via five-fold cross-validation, which increases the total computational burden dramatically without qualitatively changing the outcomes. These results are available upon request.

Table A1: Simulation Results on Risk Comparison for $T = 50$

Method	Homoskedasticity		Heteroskedasticity	
	SDFE	MAFE	SDFE	MAFE
LS	1.4546	1.1620	1.2652	0.9903
LASSO	1.4069	1.1317	1.2382	0.9870
RT	1.5026	1.1940	1.1700	0.8758
BAG	1.2579	1.0005	1.0035	0.7503
RF	1.2540	1.0007	1.0155	0.7672
SVR_{L}	1.5039	1.1947	1.3076	1.0062
SVR_{G}	1.3820	1.1062	1.1888	0.9354
LSSVR_{G}	1.2854	1.0264	1.0507	0.7977
$\text{LSSVR}_{\text{G}}^{\text{SA}}$	1.2532	1.0001	1.0350	0.7997
$\text{LSSVR}_{\text{G1}}^{\text{MA}}$	**1.2423**	**0.9908**	**0.9418**	**0.6909**
$\text{LSSVR}_{\text{G2}}^{\text{MA}}$	1.2436	0.9918	0.9528	0.7011

Note: Bold numbers denote the method with the best performance in that column of the table.

$\text{LSSVR}_{\text{G1}}^{\text{MA}}$ has the best relative performance according to all criteria, although the performance of its heteroskedasticity-robust version, $\text{LSSVR}_{\text{G2}}^{\text{MA}}$, is fairly close. Overall, we observe lower risks in the heteroskedastic scenario.

On the other hand, $\text{LSSVR}_{\text{G}}^{\text{SA}}$ exhibits less impressive performance. This finding is not surprising since the set of candidate strategies is constructed from the full permutation of all the predictors without any prior screening. Obviously, candidate models that incorporate only the irrelevant predictors, tend to generate unsatisfactory predictions. The use of equal weights in $\text{LSSVR}_{\text{G}}^{\text{SA}}$ is unable to diminish the impact of poor forecasts and thus causes the efficiency loss.

We extend the above exercise with expanding training sample sizes of $T = 50, 100, ..., 500$. The outcomes are plotted in Figure A1, in which panels (a) to (d) report the SDFE under homoskedasticity, the MAFE under homoskedasticity, the SDFE under heteroskedasticity, and the MAFE under heteroskedasticity, respectively. To avoid the figure being cluttered, we only present the results by LS, LSSVR_{G}, $\text{LSSVR}_{\text{G}}^{\text{SA}}$, and $\text{LSSVR}_{\text{G}}^{\text{MA}}$ under $C_1'(w)$, which are captured by dotted, dash-dotted, dashed, and solid lines, respectively. For presentation convenience, we standardize all results by the risk of LS. The sample size is presented on the horizontal axis and the estimated relative risk is displayed on the vertical axis.

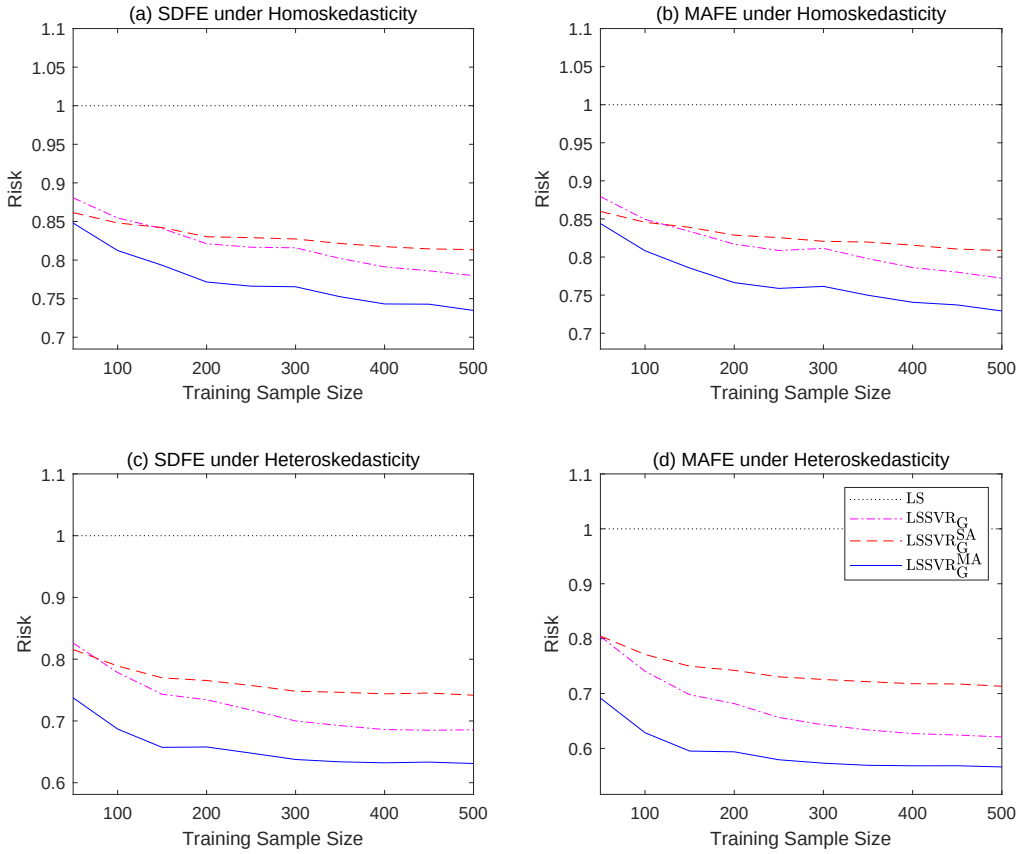
The pattern in Figure A1 is consistent. The results by LS are deliberately fixed at one for all T since it acts as the benchmark. The lines of risks by LSSVR_{G}, $\text{LSSVR}_{\text{G}}^{\text{SA}}$, and $\text{LSSVR}_{\text{G}}^{\text{MA}}$ are all downward sloping indicating that their gains relative to LS strengthen as the sample size T increases. $\text{LSSVR}_{\text{G}}^{\text{SA}}$ is outperformed by LSSVR_{G} in most cases. In contrast, $\text{LSSVR}_{\text{G}}^{\text{MA}}$ is always below LSSVR_{G} in each subplot, which verifies the advantage of $\text{LSSVR}_{\text{G}}^{\text{MA}}$ as opposed to LSSVR_{G}. We also notice that the relative risks by $\text{LSSVR}_{\text{G}}^{\text{MA}}$ are lower under heteroskedasticity than those under homoskedasticity.

It is also interesting to further investigate the improvement of MAML over its base method. As a representative case, the comparison between $\text{LSSVR}_{\text{G}}^{\text{MA}}$ and LSSVR_{G} is conducted through the computed improvement ratio (IR)

$$\text{IR} = \frac{r_{\text{LSSVR}_{\text{G}}} - r_{\text{LSSVR}_{\text{G}}^{\text{MA}}}}{r_{\text{LSSVR}_{\text{G}}^{\text{MA}}}} \times 100\%, \tag{A56}$$

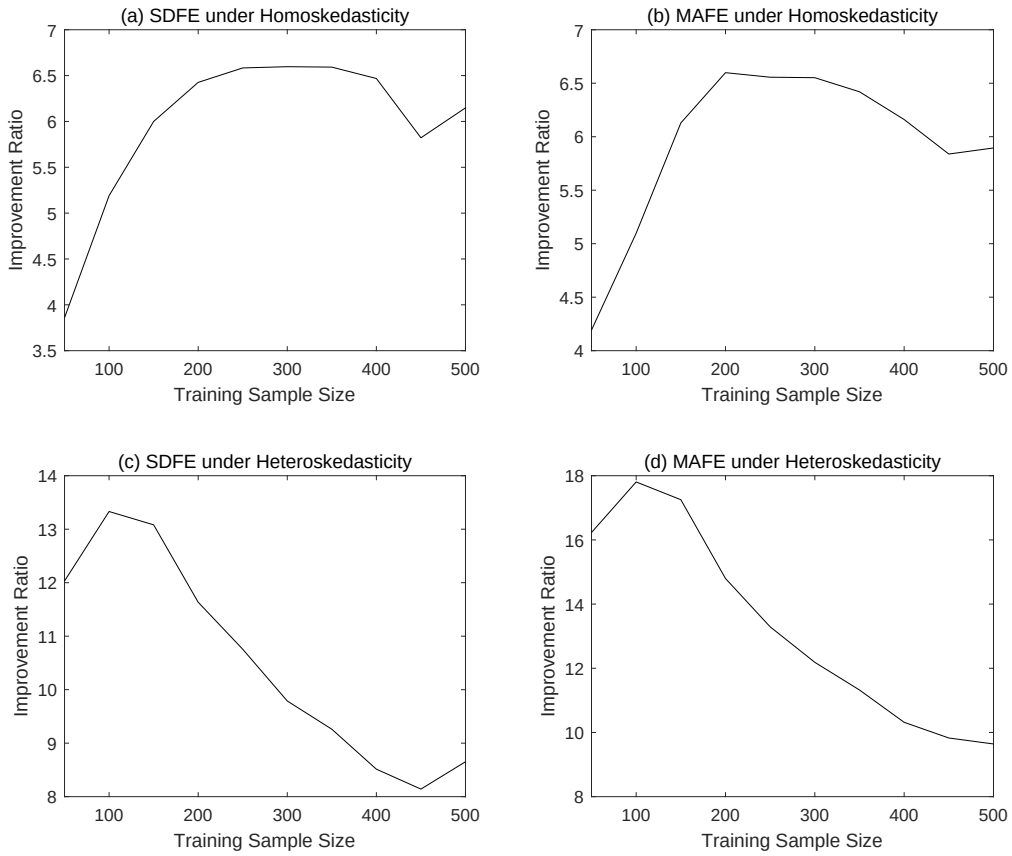
where $r_{\text{LSSVR}_{\text{G}}}$ and $r_{\text{LSSVR}_{\text{G}}^{\text{MA}}}$ denote the respective risks by LSSVR_{G} and $\text{LSSVR}_{\text{G}}^{\text{MA}}$. Results for

Figure A1: Relative Performance with Varying Sample Sizes



Notes. The figure reports the SDFE and MAFE estimates of LSSVR_{G}, $\text{LSSVR}_{\text{G}}^{\text{SA}}$, and $\text{LSSVR}_{\text{G}}^{\text{MA}}$ relative to LS estimates of a generalized unrestricted model with all the p predictors. Panels (a) and (b) present results for the homoskedasticity scenario, while panels (c) and (d) present outcomes for the heteroskedasticity scenario.

Figure A2: Improvement Ratio of $\text{LSSVR}_{\text{G}}^{\text{MA}}$ relative to LSSVR_{G}



Notes. The figure reports the risk improvement ratios of $\text{LSSVR}_{\text{G}}^{\text{MA}}$ relative to LSSVR_{G}. Panels (a) and (b) present results for the homoskedasticity scenario, while panels (c) and (d) present results for the heteroskedasticity scenario.

$T = 50, ..., 500$ are depicted in Figure A2, where panels (a) to (b) correspond to the IR ratios under homoskedasticity and panels (c) to (d) present the IR ratios under heterokedasticity. The horizontal axis represents the training sample size and the vertical axis displays the estimated improvement ratio. The outcomes further confirm that there are benefits from the model averaging estimators, with higher improvement ratios in the case of heteroskedasticity and smaller sample sizes.

E Locations of 35 Stores in China

Table A2 provides a complete list of 35 stores of the footwear brand and their locations in China as of March 25, 2022, the end of our sample period. The city with the most number of stores is Shanghai, with seven retail stores, which is about 20% of all the retail stores in China. All the stores in Shanghai are evenly spread across six districts. Chongqing, Nanjing and Beijing are the next three cities in the ranking, with a total number of 12 stores.

Table A2: Geographic Locations of 35 Stores

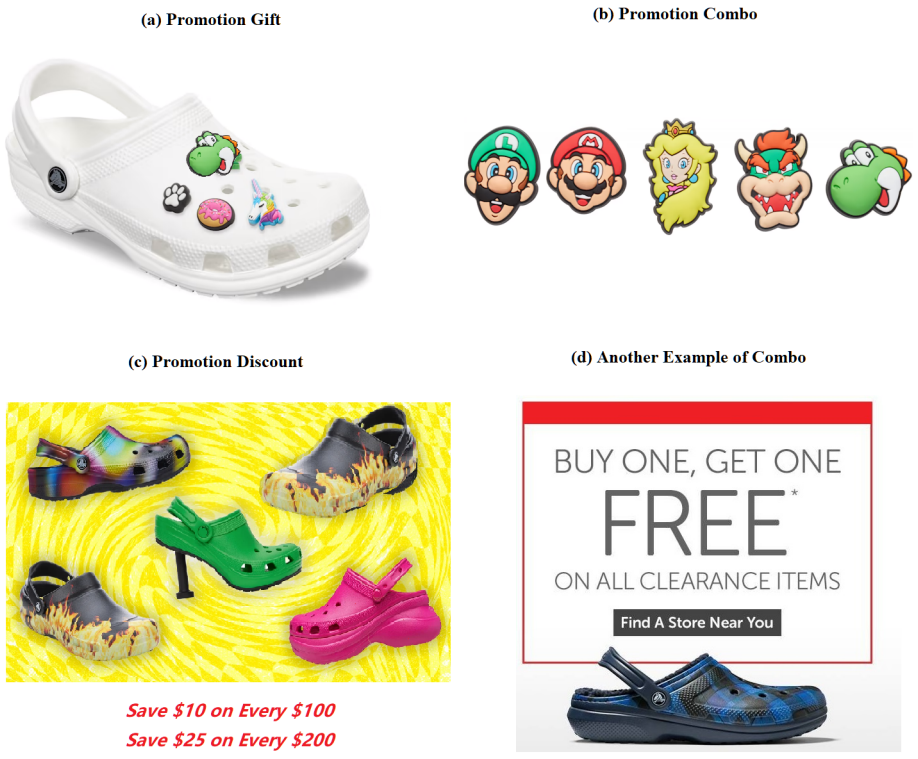
Number	City	District
1	Shanghai	Qingpu
2	Guangzhou	Zhuhai
3	Shenyang	Hunnan
4	Shanghai	Pudong
5	Xi'an	Lintong
6	Nanjing	Xixia
7	Nanjing	Jiangning
8	Wuxi	Wuxi
9	Nanjing	Jiangning
10	Chongqing	Jiulongpo
11	Xi'an	Gaoxin
12	Shijiazhuang	Luquan
13	Kunming	Guandu
14	Chongqing	Yubei
15	Chongqing	Jiangbei
16	Shanghai	Jing'an
17	Hefei	Gaoxin
18	Beijing	Chaoyang
19	Tianjin	Wuqing
20	Shanghai	Pudong
21	Wuhan	Huangpo
22	Beijing	Chaoyang
23	Xianyang	Weicheng
24	Shanghai	Putuo
25	Hangzhou	Yuhang
26	Beijing	Dongcheng
27	Chengdu	Pixian
28	Nanjing	Xuanwu
29	Chongqing	Bishan
30	Chongqing	Beibu
31	Hangzhou	Qianjiang
32	Shanghai	Yangpu
33	Shanghai	Xuhui
34	Chengdu	Jinjiang
35	Guangzhou	Tianhe

F Comprehensive Overview of Data Source and Promotion Strategies

Our study focuses on analyzing weekly sales data from *Crocs Retail, LLC (China)* between July 5, 2021, and March 25, 2022, comprising 35 stores and 38 weeks. The dataset, after cleaning, contains 1,168 observations. The primary variables of interest are the weekly number of effective in-store customers and the corresponding sales revenue at the store level. We explore factors affecting future sales, including past sales records, one-week-ahead weather forecasts at the district level, and different promotion activities. We categorize promotion strategies as (i) promotion gifts, (ii) promotion combos, and (iii) promotion discounts. While these strategies are available to all stores, shop assistants guide customers on their options, and multiple promotions can be used concurrently in a single purchase. The main objective is to understand the impact of various promotion strategies on Crocs store-level sales.

For concreteness, the four plots in Figure A3 exemplify the three promotion strategies. A gift item is usually an accessory or a tag-along that values much less than the main item. As shown in Figure A3(a), the four Jibbitz are free gifts with purchase of the clogs. Promotion combo always involves buying a bundle of items together like the five Jibbitz in Figure A3(b). The "buy one get one free" promotion in Figure A3(d) gives another example of promotion combo, where the promotion item is not an accessory and is restrictive to be of the same type as the purchased item (possibly with varying colors or sizes). Both gift and combo promotions are constrained to certain items decided by the store managers, which leaves the customers with limited choices. Promotion discount on the other hand is more flexible. Although the discount percentages are tiered by the total amount of purchase as displayed in Figure A3(c), there are usually no constraints on the applicable items.

Figure A3: Illustration of Promotion Strategies



G Quasi-Experimental Analysis of Three Promotion Predictors

In this section, we explore causality between three promotion predictors and two sales response variables using quasi-experiments derived from the original dataset. We begin by forming treated groups that consist of observations simultaneously implementing all three promotion strategies: Gift, Combo, and Discount. We then construct three untreated groups for each individual promotion strategy: Gift, Combo, and Discount. Each untreated group consists of entries that have never implemented a particular promotion strategy, but still have the history of utilizing the other two strategies. The above approach allows us to control for the influence of other promotion strategies and enables a more accurate measurement of the treatment effect from the target promotion strategy.

Due to the inherent nature of the original sample, we can treat these exercises as quasi-natural experiments. The untreated groups for Gift, Combo, and Discount strategies comprise 943, 707, and 796 observations, respectively, and each corresponding treated group consists of 669 observations. In our analysis, we employ the fixed effect regression on each sample and present the resultant treatment effect estimates in Table A3.

The outcomes imply that the estimated treatment effects on the two response variables, namely the number of effective customers and sales revenue, indicate insignificance for the Gift promotion. Conversely, they demonstrate significant positivity for the Combo promotion and significant negativity for the Discount promotion at a 1% level of significance. These findings are consistent with the results of the fixed effect regression shown in Table 2. In conclusion, the findings from this exercise establish a causal relationship between the three promotion strategies and the two sales responses, providing additional support for the subsequent evaluation of the marginal effects of these promotion strategies.

Table A3: Treatment Effect Estimates from Fixed Effect Regressions

	Gift		Combo		Discount	
	# of Customers	Sales Revenue	# of Customers	Sales Revenue	# of Customers	Sales Revenue
Treatment Effect	0.0678	0.0623	1.2429***	0.7099***	-0.2616***	-0.1596***
	(0.0973)	(0.0524)	(0.2748)	(0.1953)	(0.0594)	(0.0413)

Notes. This table presents treatment effect estimates obtained from a panel regression with fixed effects applied to the quasi-natural experiments. The numbers in parentheses represent the heteroskedasticity-robust standard errors. Additionally, superscripts *, **, and *** denote significance levels at 10%, 5%, and 1%, respectively, for the associated coefficients.

H Additional Empirical Results

We also examine the effect of alternative bootstrap on the ranking of predictor importance. Conventional bootstrap is usually executed on the cross-sectional data since it is unnecessary to consider the chronological order for each store in our panel data. Here we consider the moving block bootstrap formulated by Künsch (1989) as an alternative resampling method. Instead of performing single-data resampling, Künsch (1989) advocated the idea of resampling blocks of observations at a time. By retaining the neighboring observations together within each block, the dependence structure of the random variable at short lag distances is preserved. See Kreiss and Lahiri (2012) for a detailed literature review.

We treat the store-wise data as blocks for the block bootstrap method. In total, we have 35 blocks corresponding to the 35 stores. Table A4 present the top 10 most important predictors using two alternative bootstrap methods, where $\text{LSSVR}^{\text{MA}}_{\text{G1}}$ is employed as the trained strategy. Columns 2 to 4 replicate the results of $\text{LSSVR}^{\text{MA}}_{\text{G1}}$ from Table 5 by the regular bootstrap. Findings by block bootstrap are presented in Columns 4 to 5. The two response variables are listed in the second row for regular bootstrap and block bootstrap, respectively.

As we can see, the results by block bootstrap are virtually identical to those by regular bootstrap. The previous sales predictors are still highly important and the two promotion variables P_{Combo} and P_{Gift} remain to be more important than $\text{P}_{Discount}$ like what demonstrated in Table 5. The above finding confirms the robustness of our ranking to alternative bootstrap methods.

Table A4: Top 10 Most Important Predictors Using Alternative Bootstraps

Ranking	Regular Bootstrap		Block Bootstrap	
	# of Customers	Sales Revenue	# of Customers	Sales Revenue
1	lag(Revenue)	lag(Revenue)	lag(Revenue)	lag(Revenue)
2	lag(Customer)	lag(Unit)	lag(Customer)	lag(Unit)
3	lag(Unit)	lag(Customer)	lag(Unit)	lag(Customer)
4	P_{Gift}	P_{Combo}	P_{Combo}	Temp_{Max}
5	P_{Combo}	P_{Gift}	Temp_{Max}	P_{Combo}
6	Temp_{Max}	Temp_{Max}	P_{Gift}	P_{Gift}
7	Temp_{Avg}	Temp_{Avg}	Temp_{Avg}	Temp_{Avg}
8	$\text{P}_{Discount}$	Off Rate	Off Rate	Off Rate
9	Off Rate	$\text{P}_{Discount}$	Temp_{Min}	$\text{P}_{Discount}$
10	Temp_{Min}	Temp_{Min}	$\text{P}_{Discount}$	Temp_{Min}

Notes: The above results are formed on $\text{LSSVR}^{\text{MA}}_{\text{G1}}$.

H.1 Pooling Vs. Individual

In this section, we compare the out-of-sample performance of the pooling LS regression (denoted as LS_{pool}) with that of the store-wise individual LS regressions (denoted as $LS_{individual}$). The pooling LS assumes that all the stores share common coefficients, whereas the individual LS models acknowledge store heterogeneity by treating each store-wise subsamples independently but ignores any possible correlation among the stores.

The forecasting exercise in the main text is replicated here. We use the first 20 weeks of data as the training set and conduct one-week-ahead forecasts for each store. Note that for the individual LS regressions, the (subsample of) training set is a simple time series. We calculate SDFE and MAFE of the two approaches and present the results in Table A5. Results by LS_{pool} are identical to those from the main text.

Table A5: Forecasting Results of Sales Response Variables by Two LS Regressions

Method	SDFE		MAFE	
	# of customers	Sales revenue	# of customers	Sales revenue
LS_{pool}	**33.3456**	**15.9741**	**19.2915**	**9.1839**
$LS_{individual}$	451.3338	166.4492	82.9053	38.8010

Notes. The results of the prediction exercise for the two sales responses are reported in this table. The risks of the forecasting exercise are evaluated by the SDFE and MAFE presented in the left and right panels, respectively. Bold numbers denote the estimator with the best performance in that column of the table.

As we can see, the forecasting accuracy by LS_{pool} is much higher than $LS_{individual}$ in all cases which supports our regression setup that pools all the store-wise data together. This evidence coincides with the findings by Ali et al. (2009) that data pooling always improves model performance. This is understandable since the pooling regression improves the forecast accuracy of individual store sales by utilizing information common to all the stores. On the one hand, the poor performance of individual store-level regressions implies that ignoring the potential correlation among the stores can be quite costly for forecasting accuracy. We also note that individual LS regression may suffer from severe curse of dimensionality as the total number of observations is always no larger than 20 for each time series model.